Journal of Pathology Informatics 14 (2023) 100340

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Handling DNA malfunctions by unsupervised machine learning model

Mutaz Kh. Khazaaleh [a,*], Mohammad A. Alsharaiah [b], Wafa Alsharafat [c], Ahmad Adel Abu-Shareha [b], Feras A. Haziemeh [a], Malek M. Al-Nawashi [a], Mwaffaq abu alhija [b]

[a] Department of Computer Science, Al-Balqa Applied University, Al-Salt, Jordan
[b] Department of Data Science and Artificial Intelligence, Al-Ahliyya Amman University, Amman, Jordan
[c] Department of Information Systems, Al al-Bayt University, Mafraq, Jordan

ARTICLE INFO

Keywords:
Cell cycle
DNA damage
Cell fate
Unsupervised machine learning
K-means clustering

ABSTRACT

The cell cycle is a rich field for research, especially, the DNA damage. DNA damage, which happened naturally or as a result of environmental influences causes change in the chemical structure of DNA. The extent of DNA damage has a significant impact on the fate of the cell in later stages.

In this paper, we introduced an Unsupervised Machine learning Model for DNA Damage Diagnosis and Analysis. Mainly, we employed K-means clustering unsupervised machine learning algorithms. Unsupervised algorithms commonly draw conclusions from datasets by solely utilizing input vectors, disregarding any known or labeled outcomes. The model provided deep insight about DNA damage and exposes the protein levels for proteins when work together in sub-network model to deal with DNA damage occurrence, the unsupervised artificial model explained the sub-network biological model activities in regard to the changing in their concentrations in several clusters, they have been grouped in such as (0 - no damage, 1 - low, 2 - medium, 3 - high, and 4 - excess) DNA damage clusters.

The results provided a rational and persuasive explanation for numerous important phenomena, including the oscillation of the protein p53, in a clear and understandable manner. Which is encouraging since it demonstrates that the K-means clustering approach can be easily applied to many similar biological systems, which aids in better understanding the key dynamics of these systems.

Introduction

The cell cycle is a vital process which produces 2 cells through the division of the mother cell, also this process called replication.[4] A cell passes through 3 stages during cell division: interphase, mitosis, and cytokinesis.[2] The cell cycle has a very precise control system known as regulatory system. This system consists of 3 checkpoints: 1 - G1/S, 2 - G2/M, and 3 - M to ensure that the cell cycle in each phase has been correctly completed.[31] The cell cycle consists of a group of proteins called cyclines and cycline-dependent kinases that interact together under checkpoints supervision, even though the major actors in cell cycle regulation are p53 and RB protein.[4]

DNA damage, which can happen for a variety of reasons and has a significant impact on how the cell cycle progresses, is the main contributor. If DNA damage occurs, there are several paths a cell can pass through (damage recovery or cell death).[8,12,26] To explain G1/S checkpoint, several models have been proposed including, but not limited to models.[5,14,15,18,20,30,33–35]

For this research, we used a model proposed by Khazaaleh et al[21] for DNA damage signaling pathway to know the cell fate based on p53 oscillation. p53 protein plays a major role in triggering the control mechanisms in the cell cycle. The model is used to determine the network's structure based on the DNA damage system's chemical reactions for G1/S checkpoint as shown in Fig. 1.

Numerous protein kinases are drawn to the region of DNA damage as a result of DNA damage, and they start a signaling cascade that stops the cell cycle. Depending on the type of harm incurred, ATM/ATR is the initial kinase at the point of damage, p53 is a gene-regulatory protein that produces phosphate phosphorylation. In its natural state, Mdm2 promotes p53 ubiquitination and degradation in proteasomes by binding to it. Because p53's ability to bind to Mdm2 is inhibited by phosphorylation, p53 builds up to high levels and promotes transcription of the gene encoding the CKI protein p21. The G1/S-Cdk and S-Cdk complexes are bound and inactivated by the p21, which causes the cell to be arrested in G1.[1]

In this research, we have a large and unconnected dataset that represents the concentration of proteins during the cell cycle, DNA damaged at different levels. We used machine learning model by using K-means algorithm to organize these data and find relationships between them, to get deep understanding of the concentration of the main proteins at different levels of DNA damage.

* Corresponding author.
E-mail address: mutaz.khazaaleh@bau.edu.jo (M.K. Khazaaleh).

http://dx.doi.org/10.1016/j.jpi.2023.100340
Received 30 May 2023; Received in revised form 25 September 2023; Accepted 9 October 2023
Available online 17 October 2023

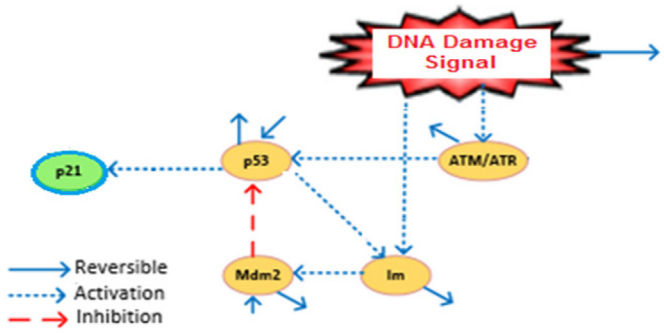

Fig. 1. The DNA damage signalling pathway by Khazaaleh et al.[21]

A better understanding of the DNA damage is useful for understanding many of the diseases and find the treatments of these diseases, such as cancer disease.

Literature review

Several machine learning techniques are used in several models of biochemical reaction network to get deep understanding for these networks. In this section, many researches of applying K-means algorithm on the cell cycle data are addressed.

Wu et al[36] presented the GKMCA genetic clustering by using K-means method for categorizing gene expression datasets. The GKMCA employs 3 operations on genetic (selection, crossover, and mutation) and an IOKM operator stemming from IOKMA. Each individual is represented by a table, independently selecting a clustering. The GKMCA demonstrates superior performance over IOKMA and other GA-clustering algorithms that do not utilize the IOKM operator when applied to 2 authentic gene expression datasets.

Duan and Zhang[10] proposed a new approach to enhance the k-means method's performance. As biological benchmarks, protein complexes from a public website were employed. By evaluating the "weighted k-means" technique on a dataset of yeast cell cycle, they applied a progressively decreasing weight on the variable level of k-means approach, they used the modified Rand index to gauge how well the k-means clusters agree with the protein complex structures in order to compare them. The final results proved that using a weight function exp[− (1/2)(t2/C2)] with C around the length of 2 cell cycles causes to a large improvement in the effectiveness of the k-means clustering.

In 2006, Chan and his team proposed a research to solve global gene trajectory clustering. This research employs a brand-new global clustering technique dubbed the Greedy Elimination Method (GEM). GEM is easy to use and works wonders to increase the solutions' overall optimality. Studies comparing the GEM to the traditional K-means and the greedy incremental technique are applied on 2 set data contains gene expressions revealing that the GEM scores much reduced grouping errors.[6] This method is simple to implement and apply, but it needs to be tested on a larger number of datasets to prove its effectiveness.

According to Sivozhelezov et al[32] study, after identifying the major genes involved in the T cells in human cell cycle, a variety of clustering methods were applied to the genes according to the stated scores. The identical 6 "leader" genes elaborate in regulation of human T lymphocytes cell cycle were consistently chosen by all clustering techniques used, including K-means and hierarchical. The 6 genes' relationships to experimental findings defining the transition of human T cells between cell cycle phases are reviewed.

Jaeger and his collages offered a method for automatically identifying cell cycle phases using 3D spinning disk confocal imaging data of embryonic fibroblast mouse cells. They segment each volume using a 3D k-means technique, and then they extract a collection of shape and curvature characteristics to describe the subcellular foci patterns connected to each channel's cell cycle phases. For 5 phases on the cell cycle, they achieve accuracy of about 92%, and their scalability is encouraging.[19]

The eXploratory K-means (XK-means) approach for grouping gene expression is introduced in Lam and Tsang[24] study. The approach was built by using incorporation of an exploratory mechanism and the K-means to avoid the clustering process convergent too early. According to experimental findings, XK-means outperforms existing evolutionary algorithm-based techniques in terms of speed, inaccuracy, and stability when grouping gene expressions. The suggested solution is less sophisticated and is simple to apply in real-world situations.

In order to identify stained nuclei and separate them for analysis of DNA content at various cell cycle phases, Ferro et al presented a novel fluorescence image-based framework. The methodology involves using discriminating characteristics, such as area and total intensity, acquired using fluorescence microscopy from in-situ labeled nuclei. This enables the evaluation of the cell cycle phase of individual cells and subpopulations. The Gaussian mixture model classification system is used to improve analytical frameworks, and it allows for very precise classification clusters according to phases: 1 - G1, 2 - S, and 3 - G2. The results show that the imaging framework is strong at recognizing specific DAPI-colored nuclei and determining their proper cell cycle phases.[11]

Recently, in 2023 many of research has been published in biological applications based on machine learning.[3,7,9,13,16,17,22,27,29]

Methodology

Dataset and model

As mentioned in the Introduction section, to identify the network's structure based on the chemical processes of the DNA damage system and get dataset, we utilized a base model for DNA damage signaling pathway by Khazaaleh et al[21] model. The dataset representing the concentration values for 5 proteins during G1/S checkpoint period, resulting from applying the model at 5 levels of DNA damage.

K-means clustering involves partitioning a set of data points into groups (clusters) based on their similarity. Each group is represented by its centroid or its mean vector. This method is often used in signal processing, image compression, and data mining, involves dividing n observations into k clusters based on proximity to the mean (also known as the cluster centroid or cluster center), with each observation acting as a prototype for the cluster. James MacQueen was the first researcher to use k-means in 1967. K-means clustering is an algorithm for unsupervised learning.[28]

The K-means algorithm is employed to cluster data by partitioning samples into k groups. This technique minimizes a criterion called inertia, also known as the within-cluster variance sum-of-squares as Eq. (1).

$$\arg\min \sum_{i=1}^{K} \Sigma_{x \in Si} ||x - Mi|| \, 2 \tag{1}$$

Where:
S: sets of observations
K: number of sets of predicators
x: observation data point
M_i*: mean of points in* S_i

In order to process learning data, the K-means algorithm starts with the first randomly selected cluster that is used as a starting point for each cluster, and then performs repetition calculations to optimize the position of the cluster, and stops the creation and optimization of the cluster when either: the center is stabilized and its value is unchanged because the cluster is successful. The defined number of iterations has been achieved. As shown in Fig. 2, the dataset used as input for this model contains 3198 instances for 5 proteins, after processing in the K-means clustering model the instances are categorized into 5 clusters as output.

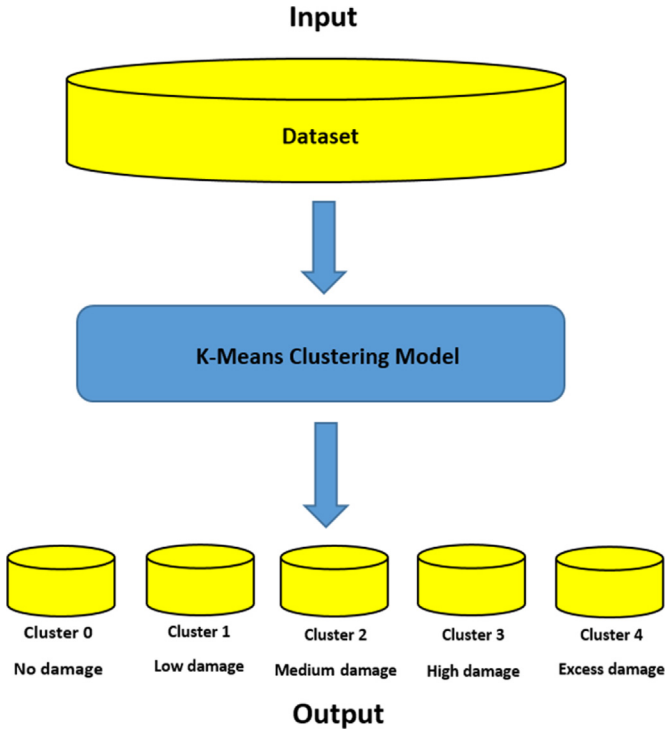

Fig. 2. The proposed K-means clustering model framework.

Simulation

- 3198 instances.
- Five levels of DNA damage: 0 - No-damage, 1 - Low, 2 - Medium, 3 - High, and 4 - Excess damage.
- 25 attributes as following:
  - p21 with 5 DNA damage levels.
  - p53 with 5 DNA damage levels.
  - Mdm2 with 5 DNA damage levels.
  - ATM/ATR with 5 DNA damage levels.
  - Im with 5 DNA damage levels.
- Number of iterations: 20.

Results and discussion

Table 1 shows the final cluster centroids and instances percent for each cluster.

As shown in Table 1 and Fig. 3, we find that only 4% of instances are in cluster 0 (No damage) and this is caused by most of the proteins are stay in stead and low level when no DNA damage has occurred. We also note that the instances rate begins to rise to become 23%, 7%, and 17% for cluster 1 (low damage), cluster 2 (medium damage), and cluster 3 (high damage), respectively. DNA damage results in p53 activation, which triggers p21. The role of p21 is to prevent CDK from acting in order to cause cell cycle arrest by preventing the phosphorylation of Rb and keeping E2F inactive.[37,38] This gives an explanation of the increase in the number of instances in cluster 1 (low damage). In medium damage, more activation occurs for p53 that

Table 1
Final cluster centroids.

	Full data	Cluster#				
		0	1	2	3	4
Attribute	(3198.0)	(129.0)	(738.0)	(219.0)	(542.0)	(1570.0)
Instances percent	100%	4%	23%	7%	17%	49%

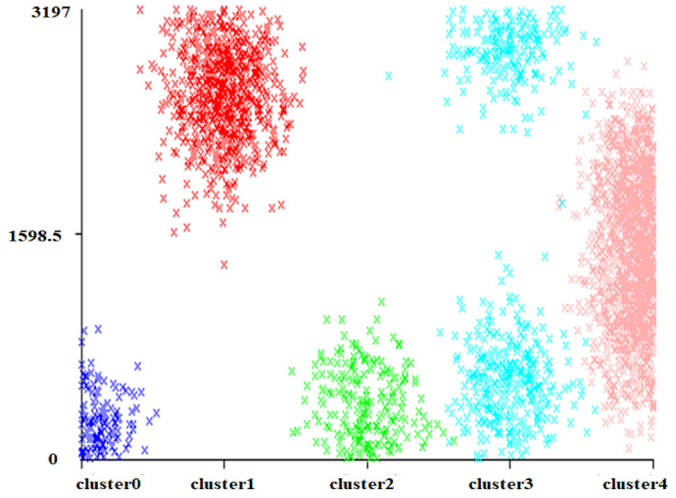

Fig. 3. 3198 instances for 5 protiens in 5 DNA damage clusters: no damage, low, medium, high, and excess DNA damage.

cause large induces p21. Once DNA damage is repaired, the negative feedback loops involving Mdm2 and p53 can be reestablished completely, causing p53 levels to revert to a low state. Consequently, the reduction in p53 results in decreased p21 levels, accounting for the fewer occurrences observed in cluster 2 (medium damage) and cluster 3 (high damage) compared to cluster 1 (low damage).

In the process of regulating excessive DNA damage, the sequential activation of p53 and Mdm2 occurs, leading to elevated protein concentration levels, which explains the high number of instances in cluster 4 (excess damage) 49%. Because of the inability to DNA recovery process, the cell goes on a path of terminating itself. All of what we mentioned earlier which matched the results of the experiments conducted by Lahav et al, Lev Bar-Or et al, and Yu et al.[23,25,37]

Fig. 4 shows the instances for 5 protiens (1 - p21, 2 - p53, 3 - Mdm2, 4 - ATM/ATR, and 5 - Im) in 5 DNA damage clusters: A. No damage, B. Low, C. Medium, D. High, and E. Excess DNA damage.

As shown in Fig. 5A, which represents p21 concentration with no DNA damage, we find that most of the instances are distributed over the different clusters with a low concentration ranging between 0.000077 and 0.0079, and this is due to the fact that protein 21 does not play any role in the case of no DNA damage. Also, as shown in Fig. 5B, which represents a p21 concentration with excess DNA damage, we find that most of the instances are distributed over the different clusters with a high concentration ranging between 0.053 and 0.11, and this is due to the fact that protein 21 plays a main role in the case of excess DNA damage.

As shown in Fig. 6, which represents p53 concentration with the excess DNA damage, we note that all 5 clusters contain highly concentrated instances and at the same time contain low concentration instances, and this explains the phenomenon of oscillation in the concentration of p53 protein when the excess DNA damage occurs.

Through the final results of the proposed model and comparing them with the results of Iwamoto et al, Lahav et al, Lev Bar-Or et al, and Yu et al,[18,23,25,37] we found a significant match which confirms the effectiveness of the proposed model for dealing with big data.

Conclusion

The importance of using machine learning, especially K-means clustering method with biological systems provides a deep understanding for these systems, especially when dealing with a large dataset. In this research, we used K-means clustering technique to handle a large set of complex and unorganized data that represents the concentration of the most important proteins that are essential in the different levels of the DNA damage. The importance of the proposed model is demonstrated by achieving deeper

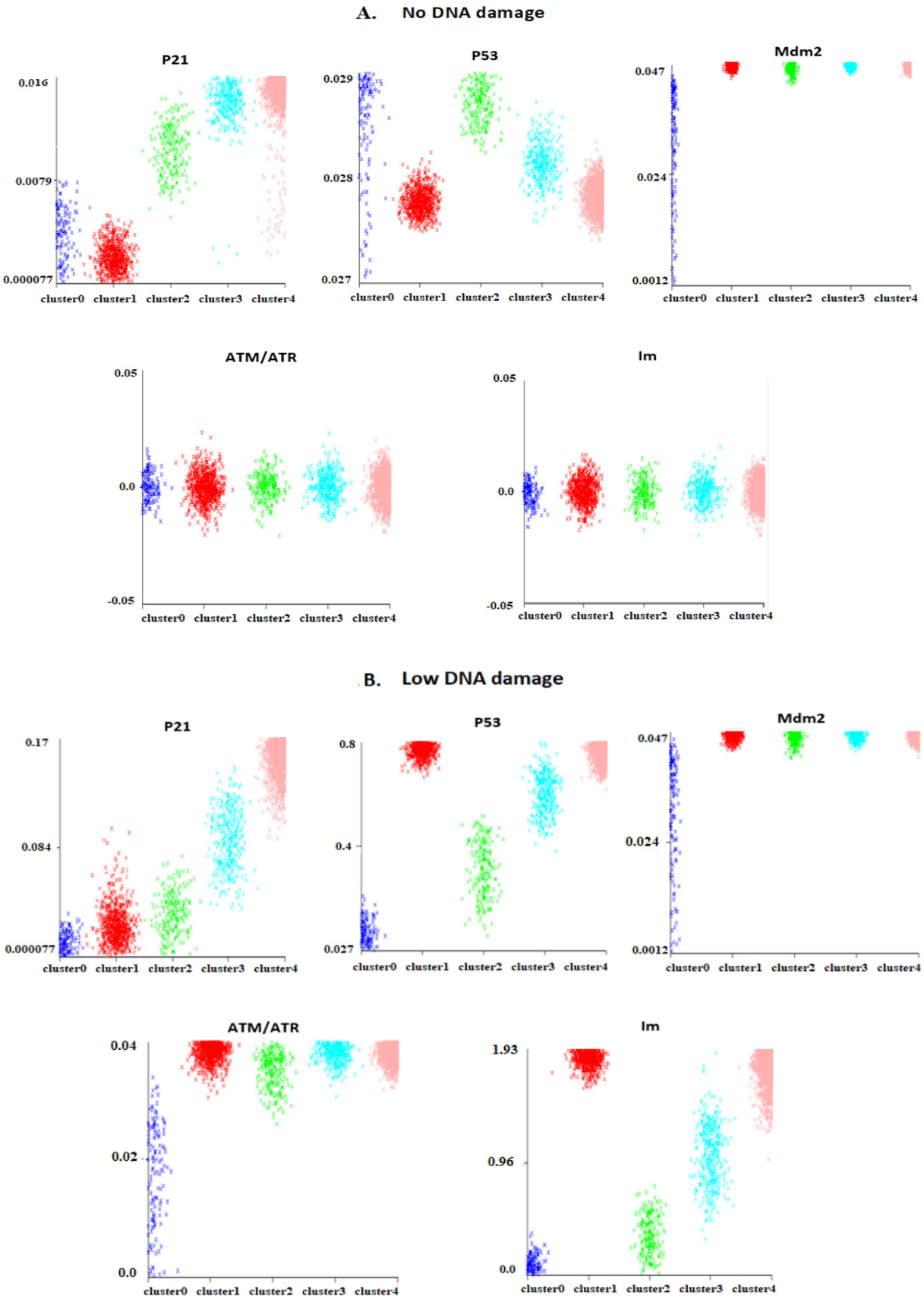

Fig. 4. Instances for 5 protiens in 5 DNA damage clusters: A. No damage, B. Low, C. Medium, D. High, and E. Excess DNA damage.

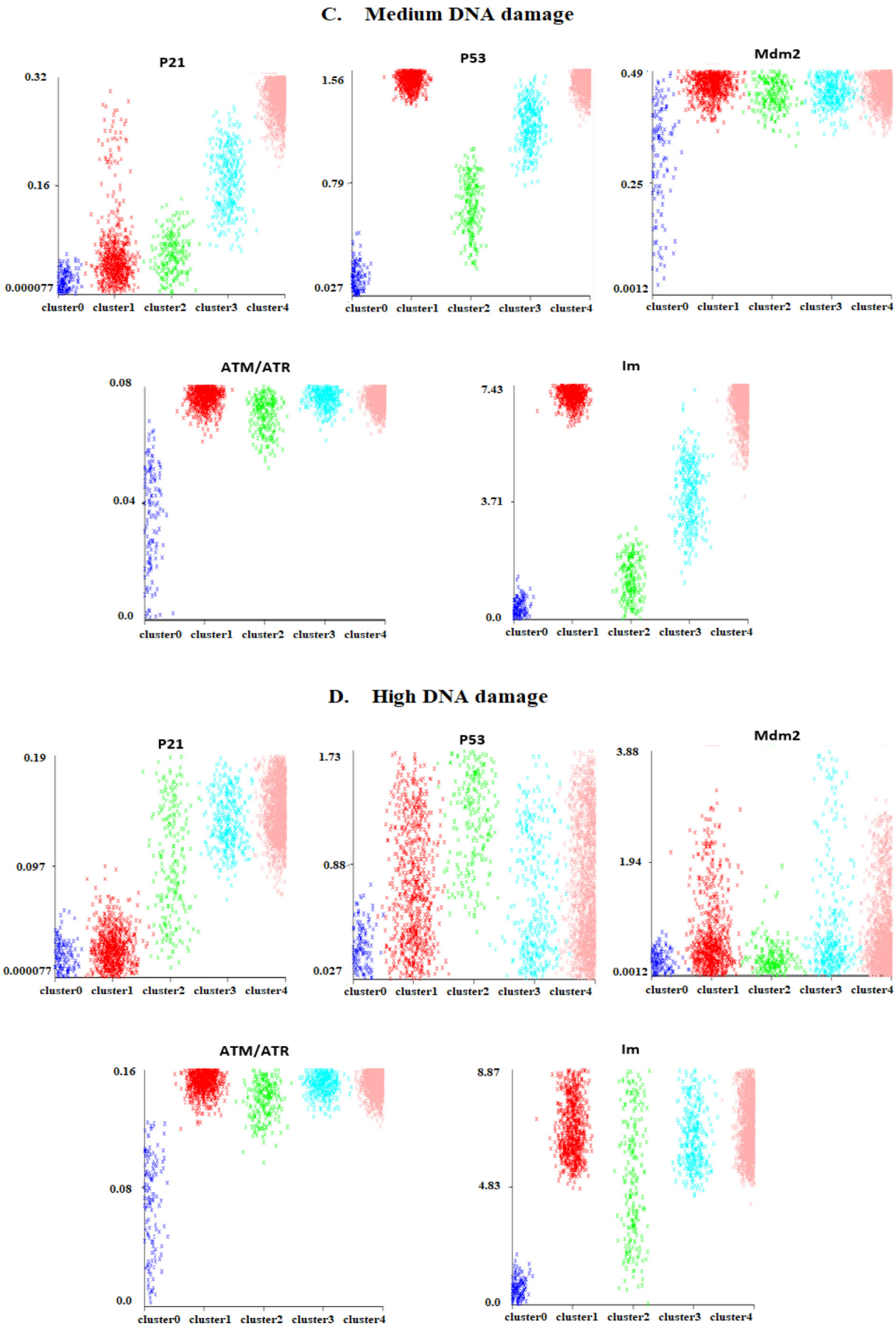

Fig. 4 (*continued*).

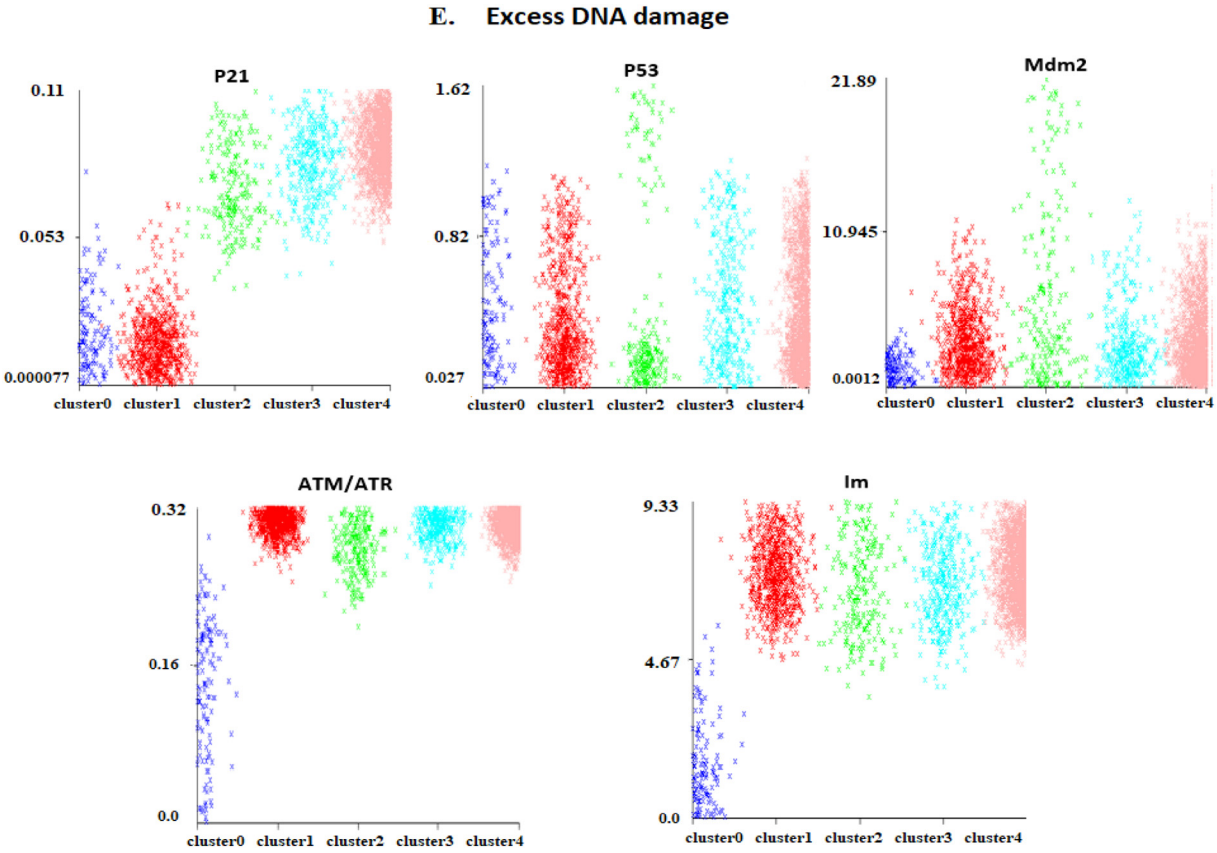

Fig. 4 (*continued*).

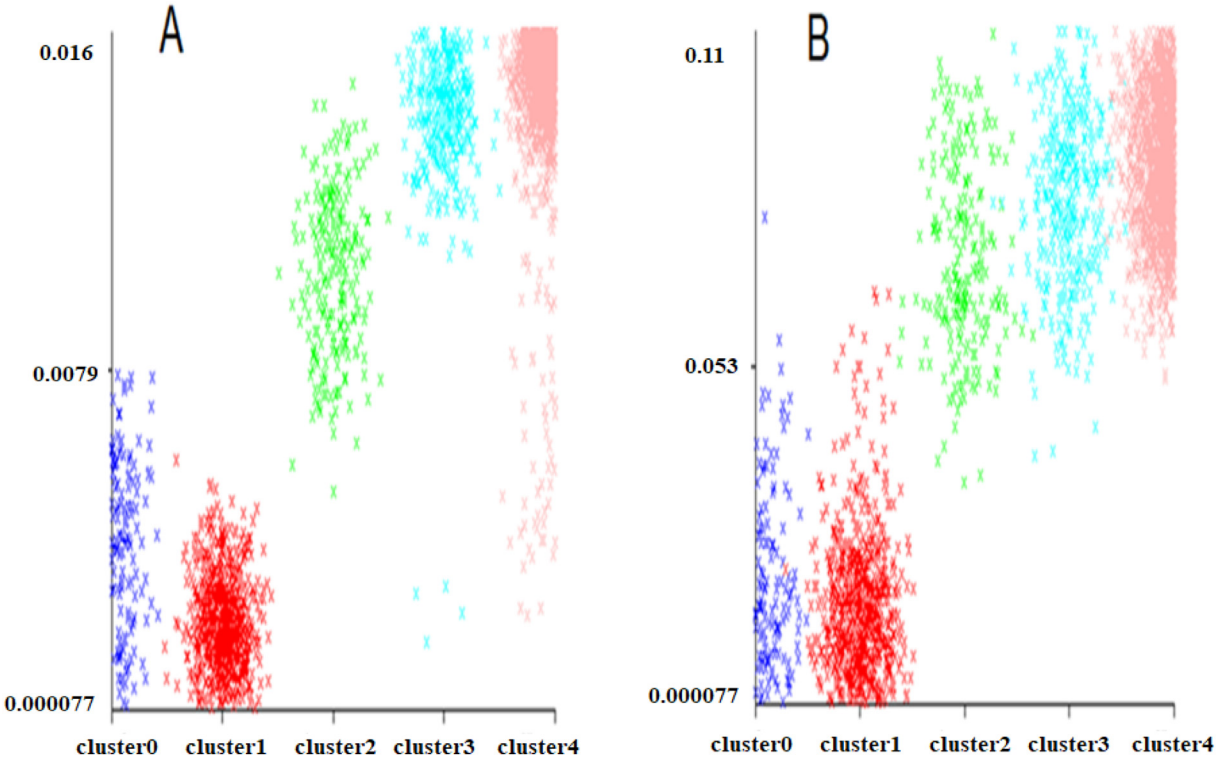

Fig. 5. Instances for p21 protien in 5 DNA damage clusters: A. No damage, B. Excess damage.

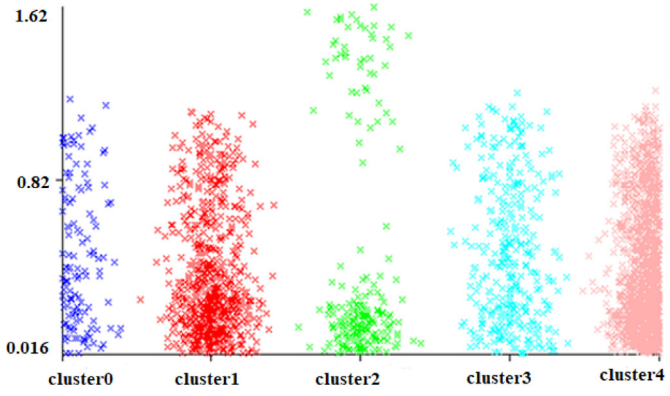

Fig. 6. Instances for p53 protien in 5 DNA damage clusters with excess damage.

understanding of the DNA damage, which means deeper understanding of numerous diseases and find cures for them, including cancer.

The results showed, in a simple and logical way, a convincing and logical explanation for many of the vital observations, such as the oscillation of protein p53. Which proves promising that the K-means clustering method can be easily applied to many similar biological systems, which helps to understand the vital dynamics of these systems in a deeper way.

In future work, we intend to improve the proposed method and to apply the proposed method on the DNA damage signaling pathway and whole cell cycle regulation and explored the effect of p53 on cell fate selection.

Funding Statement

The authors received no specific funding for this study.

Conflict of interest

The authors declare that there is no conflict regarding the publication of this paper.

References

1. Alberts B, Bray D, Hopkin K, et al. *Essential Cell Biology*. Garland Science. 2013.
2. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell.* New York: Garland Science. 2010.2008. Classic textbook now in its 5th Edition.
3. Banerjee A, Saha S, Tvedt NC, Yang LW, Bahar I. Mutually beneficial confluence of structure-based modeling of protein dynamics and machine learning methods. Curr Opin Struct Biol 2023;78, 102517.
4. Behl C, Ziegler C. *Cell Aging: Molecular Mechanisms and Implications for Disease.* Berlin: Springer. 2014.
5. Cetin NI, Bashirov R, Tüzmen S. Petri net based modelling and simulation of p16-Cdk4/6- Rb pathway. Proceedings of CEUR Workshop, Vol. 988; 2013. p. 30–44.
6. Chan ZS, Collins L, Kasabov N. An efficient greedy k-means algorithm for global gene trajectory clustering. Expert Syst Appl 2006;30(1):137–141.
7. Chandra A, Tünnermann L, Löfstedt T, Gratz R. Transformer-based deep learning for predicting protein properties in the life sciences. eLife 2023;12, e82819.
8. Ciliberto A, Novak B, Tyson J. Steady states and oscillations in the p53/Mdm2 network. Cell Cycle 2005;4(3):488–493.
9. Ding Y, Sun Y, Liu C, Jiang QY, Chen F, Cao Y. SeRS-based biosensors combined with machine learning for medical application. ChemistryOpen 2023;12(1), e202200192.
10. Duan F, Zhang H. Correcting the loss of cell-cycle synchrony in clustering analysis of microarray data using weights. Bioinformatics 2004;20(11):1766–1771.
11. Ferro A, Mestre T, Carneiro P, Sahumbaiev I, Seruca R, Sanches JM. Blue intensity matters for cell cycle profiling in fluorescence DAPI-stained images. Lab Investig 2017;97(5): 615–625.
12. Geva-Zatorsky N, Rosenfeld N, Itzkovitz S, et al. Oscillations and variability in the p53 system. Mol Syst Biol 2006;2.
13. Guo T, Li X. Machine learning for predicting phenotype from genotype and environment. Curr Opin Biotechnol 2023;79, 102853.
14. Haberichter T, Mädge B, Christopher RA, et al. A systems biology dynamical model of mammalian G1 cell cycle progression. Mol Syst Biol 2007;3(1):149–166.
15. Hatzimanikatis V, Lee KH, Renner WA, Bailey JE. A mathematical model for the G1/S transition of the mammalian cell cycle. Biotechnol Lett 1995;17(7):669–674.
16. Hu X, Fernie AR, Yan J. Deep learning in regulatory genomics: from identification to design. Curr Opin Biotechnol 2023;79, 102887.
17. Isert C, Atz K, Schneider G. Structure-based drug design with geometric deep learning. Curr Opin Struct Biol 2023;79, 102548.
18. Iwamoto K, Hamada H, Eguchi Y, Okamoto M. Mathematical modeling of cell cycle regulation in response to DNA damage: exploring mechanisms of cell-fate determination. Biosystems 2011;103(3):384–391.
19. Jaeger S, Palaniappan K, Casas-Delucchi CS, Cardoso MC. Classification of cell cycle phases in 3D confocal microscopy using PCNA and chromocenter features. Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing; 2010, December. p. 412–418.
20. Khazaaleh M, Samarasinghe S. Using activity time windows and logical representation to reduce the complexity of biological network models: G1/S checkpoint pathway with DNA damage. Biosystems 2020;191, 104128.
21. Khazaaleh M, Samarasinghe S, Kulasiri D. A new hierarchical approach to multi-level model abstraction for simplifying ODE models of biological networks and a case study: the G1/S Checkpoint/DNA damage signalling pathways of mammalian cell cycle. Biosystems 2021;203, 104374.
22. Krentzel D, Shorte SL, Zimmer C. Deep learning in image-based phenotypic drug discovery. Trends Cell Biol 2023;33(7):538–554.
23. Lahav G, Rosenfeld N, Sigal A, Geva-Zatorsky N, Levine AJ. Dynamics of the p53-Mdm2 feedback loop in individual cells. Nat Genet 2004;36:147–150.
24. Lam YK, Tsang PW. eXploratory K-means: a new simple and efficient algorithm for gene clustering. Appl Soft Comput 2012;12(3):1149–1157.
25. Lev Bar-Or R, Maya R, Lee AS, Uri A, Arnold JL, Moshe O. Generation of oscillations by the p53-Mdm2 feedback loop: a theoretical and experimental study. Proc Natl Acad Sci USA 2000;97:11250–11255.
26. Li G, Ho V. p53-dependent DNA repair and apoptosis respond differently to high-and low dose ultraviolet radiation. Brit J Dermatol 1998;139(1):3-10.
27. Li Z, Gao E, Zhou J, Han W, Xu X, Gao X. Applications of deep learning in understanding gene regulation. Cell Rep Methods 2023;100384.
28. MacQueen JB. Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press; 1967. p. 281–297.
29. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet 2023;24(2):125–137.
30. Obeyesekere MN, Zimmerman SO, Tecarro ES, Auchmuty G. A model of cell cycle behaviour dominated by kinetics of a pathway stimulated by growth factors. Bull Math Biol 1999;61(5):917–934.
31. Saltsman K. *Cellular Reproduction: Multipication by Division. Inside the Cell.* National Institute of General Medical Sciences. NIH Publication. 2005.
32. Sivozhelezov V, Giacomelli L, Tripathi S, Nicolini C. Gene expression in the cell cycle of human T lymphocytes: I. Predicted gene and protein networks. J Cell Biochem 2006;97 (5):1137–1150.
33. Sugii M, Wingender E, Matsuno H. Petri net modelling of oscillatory processes in the activation of cell cycle proteins. Proceedings of International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2013), Yeosu, Korea; 2013.
34. Tashima Y, Hamada H, Okamoto M, Hanai T. Prediction of key factors controlling G1/S phase in the mammalian cell cycle using system analysis. J Biosci Bioeng 2008;106(4): 368–374.
35. Tashima Y, Hanai T, Hamada H, Okamoto M. Simulation for detailed mathematical model of G1-to-S cell cycle phase transition. Genome Inform 2004;9:607–608.
36. Wu FX, Zhang WJ, Kusalik AJ. A genetic K-means clustering algorithm applied to gene expression data. Conference of the Canadian Society for Computational Studies of Intelligence. Berlin, Heidelberg: Springer; 2003, June. p. 520–526.
37. Yu J, Zhang L, Hwang PM, Rago C, Kinzler KW, Vogelstein B. Identification and classification of p53-regulated genes. Proc Natl Acad Sci USA 1999;96(25):14517–14522.
38. Campisi, Judith, and Fabrizio d'Adda di Fagagna. "Cellular senescence: when bad things happen to good cells." *Nature reviews Molecular cell biology* 8.9 (2007): 729-740.