DETECTION OF SUMOYLATION SITES THAT EMERGE THROUGH MUTATIONS IN CANCER

Suleyman Onur Dogan onur.dogan@std.antalya.edu.tr
Computer Science Engineering, Junior

Ecem Erdogan ecem.erdogan@metu.edu.tr
Molecular Biology and Genetics, Senior

Oznur Tastan
Computer Science and Engineering

Abstract

Post-translational modification is a general term encompassing all the chemical modifications done on proteins after translation. SUMOylation is one of the PTMs that plays a critical role in cell homeostasis and how cell responses to stress conditions. In this project we examine the mutation data acquired from cancer patients and operate our developed deep-neural network tool to predict if the mutations lead to a novel SUMOylation site. We aim to find whether the predicted SUMOylation sites are recurring in different patients and cancer types to better understand the relationship between a dysregulated SUMO mechanism with cancer development and progression.

Keywords: Sumoylation, Deep Learning, Mutation, Cancer, PTMs

1 Introduction

Post-translational modifications are the process of chemically modifying the protein after they are done being translated. PTMs include the covalent addition of subunits such as other proteins, lipids, and sugars, as well as the proteolytic removal of subunits. Overall, all the modifications regulate the activity, functionality, and localization of the proteins. Post-translational modifications occur at specific amino acid residues that are recognized and modified by corresponding enzymes. Moreover, the modifications can be reversible or irreversible, and can be reversed by another set of proteins removing the additional functional group. Since post-translational modifications play a major role in the functional proteomics regulating the cellular processes in the cell, scientists have always been giving utmost attention to the characterization of PTMs and their consequences. (Walsch, 2006)

Our project's main focus is the SUMOylation post-translational modification. SUMOylation comprises the covalent and reversible attachment of Small-Ubiquitin like Modifier (SUMO) protein to Lysine (K) residues in particular proteins. (Wilkinson et. al., 2010) There are repetitive residues characterized as

SUMO-interacting motifs, depicted as “ΨKxE”. Lysine residue positioned in the middle, Ψ symbol is standing for a hydrophobic residue and x symbol representing any amino acid residue. These consensus sites are found to be a recognition and binding site for Ubc9 conjugation enzymes however the presence of a consensus site does not guarantee the SUMOylation. Nevertheless, the motifs should be kept in consideration. (Beauclair, 2015)

SUMOylation like any other PTMs regulate and alter protein functionality and thus play a key role in many cellular downstream pathways such as transcription regulation, chromatin remodeling, DNA repair mechanism and cell cycle progression control. (Seeler et al., 2017) If there were to have a genetic mutation in the protein coding region resulting in a nucleotide change, the protein that is transcribed from that mutated nucleotide sequence region could have novel residues that should not be in that position normally. The abnormal protein sequence may eliminate or generate Lysine residues that subsequently alter the protein's normal SUMOylation state. As SUMOylation is critical for cell homeostasis mechanisms, the overexpressed or repressed SUMO cycle may potentially lead to disease states such as tumor cells due to aberrant functionality of SUMOylated proteins. Taking cancer patient data to analyze such mutations with new Lysine residues and predicting the SUMOylation state of that particular case can be a base start for cancer studies. With enough data from many cancer types and cancer patients we may be able to link the cancer state and progression with SUMOylation. Therefore, in our project we have aimed to develop a computational pipeline that will predict whether the new emerging SUMOylation sites through mutations in cancer patients are indeed SUMOylated.

2 Materials and Resources

The prediction of possible SUMO sites was carried out by “SUMOnet". SUMOnet is a pre-trained special deep neural network that does the prediction of SUMOylation based on the given 21 long amino acid sequence with Lysine residue residing in the middle, 10 amino acids before Lysine and 10 amino acids after Lysine. SUMOnet was initially designed in 3 different neural network architectures as SUMOnet 1,2 and 3. Each SUMOnet version was designed with different encodings which are One-hot, NLF and Blossum62. Among all the different SUMOnet architectures we have utilized SUMOnet-3 since, its performance is superior to other SUMonet architectures and methods available in the literature.

In the figure below the architecture of SUMOnet-3 can be seen. It is trained by Blossum62 encoded vectors, moreover, consisting of a compilation of different neural networks such as CNN, BiGRU,dense and Pooling layers. Also, figure 2 shows the evaluation metrics results of SUMOnet-3.

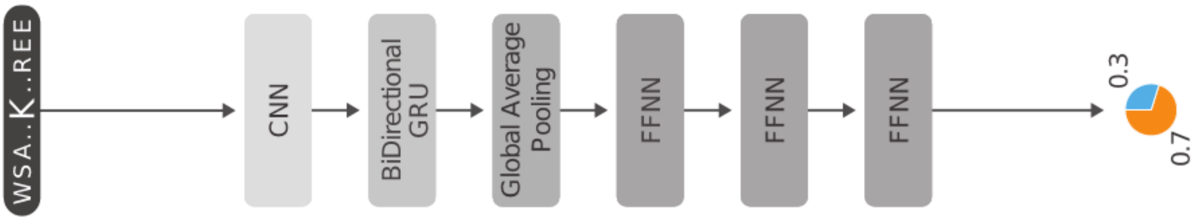
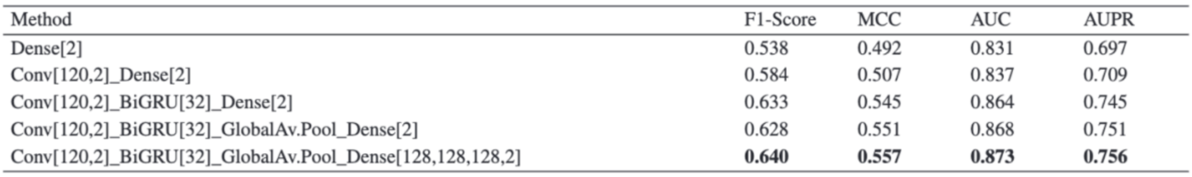


Figure 1. Architecture of deep neural network of SUMOnet-3

Combination of CNN and BiGRU layers made utility on the architecture. Therefore, one of the main parts of the SUMOnet-3 architecture is BiDirectional GRU which is able to grasp neighboring relations. Neighbor relation is a significant advantage for predicting SUMOylation because we know that nearby relationships of amino acids are significant. Even though we are focusing on the mutated lysine as a target residue, we have considered the mutation near the target residue which we will explain in chapter 3 in more detail.



Method	F1-Score	MCC	AUC	AUPR
Dense[2]	0.538	0.492	0.831	0.697
Conv[120,2]_Dense[2]	0.584	0.507	0.837	0.709
Conv[120,2]_BiGRU[32]_Dense[2]	0.633	0.545	0.864	0.745
Conv[120,2]_BiGRU[32]_GlobalAv.Pool_Dense[2]	0.628	0.551	0.868	0.751
Conv[120,2]_BiGRU[32]_GlobalAv.Pool_Dense[128,128,128,2]	**0.640**	**0.557**	**0.873**	**0.756**

Figure 2 Evaluation metrics results of SUMOnet-3

As mentioned, SUMOnet-3 requires a peptide sequence that is 21 amino acids long, with Lysine residue in the middle. In order to acquire that information, we have employed “The Cancer Genome Atlas GDC data portal”. GDC is an open-source, data-driven database where researchers can download and analyze cancer data for their studies. This platform includes cancer primary sites, different cases and projects and mutation data on each patient case. The initial step of our research was to get mutation data on Lung Adenocarcinoma (LUAD) cancer patients.

Another online resource we have used is the UniProt database. UniProt is again an open-access database, including a vast amount of protein sequence information. Genome sequencing data comprises many of the entries. Overall protein sequence, functionality, localization and even structure information is included in this database that are rendered from research. Through UniProt we have acquired the wild-type amino acid sequences of the proteins we have selected from cancer patients.

3 Methods

The main goal of the project and the tools to be used for this goal, GDC, Uniprot, SUMOnet, were mentioned above. Figure 3 shows the summarized workflow of how we can predict SUMOylation sites using these tools. The workflow consists of 4 main points, retrieving patient mutation data of particular cancer project from GDC and preprocessing the data, extracting the protein sequence from Uniprot, mapping the mutation data that comes from GDC to protein sequence and predicting possible SUMOylation on mutated lysines by using SUMOnet. Throughout the project, we wanted to create these as a tool to look at different cancer projects, and that's why all the analyzes were created to be automated. Even if we use different programming languages in some parts of the analysis, we have combined them by writing bash scripting. The bash scripting we have written takes 2 inputs, which cancer project you want to look at and where you want to download the data. All methods explained in the below are available in github as an open source project named tlmsa with functions (https://github.com/sonurdogan/tlmsa) .

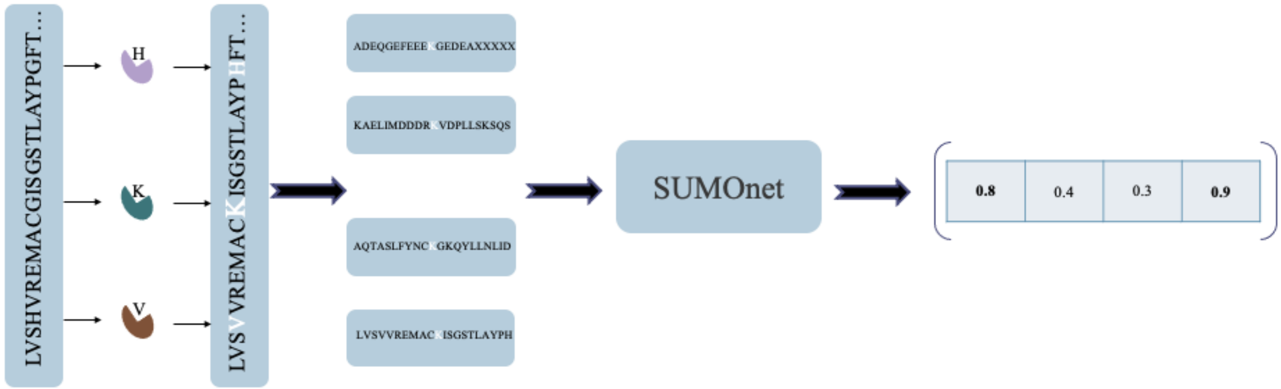


Figure 3 Summarized Workflow

3.1 Retrieving patient mutation data from GDC and Preprocessing

The first part is getting patient mutation data from GDC. As mentioned above, GDC is a database for cancer research so GDC has been used to have cancer data in our project. Since, one of our goals is creating an automated system, we have started to find a way that we can retrieve data from GDC with script and we have found the TCGAbiolinks package written in R language. Therefore, Cancer mutation data started to be obtained in R using the GDCquery function of TCGAbiolinks package. As a result of this approach, we collected patient mutation data with 140 columns for the cancer type we want, and we began filtering the columns that had essential information for us. Since we focus on mutations in the protein sequence in our project, we need columns such as patient id, mutation information, mutation position information, gene name, uniprot id. However, some of this information is not directly provided in the data so we have done feature extraction from existing features. For example, data provides us with something called HGVSp_Short column that has a unique presentation of mutation with position such as, p.N301K, we extracted amino acid and position information from the unique presentation.

After selecting the required columns and completing some feature extraction, mutation-patient filtering operations are initiated. We proceeded by selecting out mutations that resulted in lysine since SUMOylation occurs in the lysine residue and we were interested in mutations that resulted in lysine. However, mutations near the mutated K may affect SUMOylation, so we take a step back and choose all of the mutations of the genes that have a mutation resulting in K. Neighboring relationships are taken into account for the prediction outcomes at this point, as indicated in Chapter 2. The code for the data manipulation methods indicated in this part is available as retrieveData.R.

We have all the mutations and their position in the genes of all patients who have the mutation that resulted in lysine.

3.2 Extracting the protein sequences from UniProt

The mutation data was prepared after preprocessing processes, the healthy protein sequences of the genes in the data were taken from Uniprot, we looked for a program to do this automatically, and we began using the bioservices package written in Python and indirectly shifted to the Python language. At this point, the fact that we have switched to the Python programming language can be regarded a benefit because SUMOnet is a tool written in Python.

The Uniprot function of the bioservices package obtains the Uniprot ID of each gene and the sequence of the following gene; however, there may be some mutations with the same Uniprot ID, therefore, we addressed this circumstance while developing the code and optimal code in terms of time and code did not run more than one for the same gene.

3.3 Mapping mutations

We have information from two separate sources: the mutation and the wild type sequence data. We mapped the mutations to the wild type sequence at this point to obtain the patient's mutant sequence.

This part was the most challenging part in terms of coding. In order to overcome this challenging part, we created a function as well as considering the run time. Before getting into function, In our data each row is mutation with different case id, mutation position etc.

In this function, we get all the unique patient ids in the data and get each patient individually with a for loop. We get unique genes of that patient in each patient ID and again we consider each gene of that patient one by one with a for loop. If there is more than one row for that gene, it means there is more than one mutation in that gene, so we take the sequence from the first mutation for each gene and map the other mutations and the mutation in that row into a single sequence and then assign it to the other mutations/rows. At the end, we get a mutated sequence of each patient’s gene.

3.4 Possible SUMOylation site prediction by SUMOnet

Since SUMOnet's input shape is a 21-long subsequence, our goal is to focus on the mutations resulting in lysines. Therefore, before starting using SUMOnet to predict possible SUMOylation sites, input for SUMOnet has been generated with mutated Lysine residue in the middle, 10 amino acids before mutated Lysine and 10 amino acids before mutated Lysine from newly mapped mutated peptide.

We obtained the site of the mutant lysine from a newly mapped mutated peptide sequence and then constructed a 21-long subsequence near the mutated lysines. However, there were some special cases in this process, and it was that there may not be 10 amino acids before or there may not be 10 amino acids after the mutated lysine so, while writing code for generating subsequence, these cases were considered and these possible empty parts has been filled with X.

After completing all essential operations, we performed the SUMOnet-3 to predict potential sumoylation sites on different cancers and we got possible SUMOylation sites.

In addition, to gain a different perspective on our results, we explored existing motifs in the literature on the subsequence that we create. All of SUMOnet-3's prediction findings and motif results on different cancers are detailed in further detail in the Results section.

4 Results

4.1 LUAD

The initial project was to predict the SUMOylation sites on LUAD cancer patients. We have focused on primary tumor data of the LUAD patients and SUMOnet was run on 521 patients. Among those patients 6661 peptide subsequence was a candidate for a possible SUMOylation site; they all had mutated Lysine in the middle. Out of these 6661 subsequences, we have filtered based on different thresholds: 0.5, 0.7 and 0.9. The table below summarizes the positive possible SUMOylation site results as well as the presence of a consensus SUMO interacting motif within the SUMOylation site. The percentage of SUMOylation sites with the motifs are increasing as the possibility of the prediction increases.

Threshold	# of predicted SUMO sites	# of including SUMO motifs
0.5	1059	349
0.7	514	230
0.9	191	111

Table 4. Possible SUMOylation sites predictions in different thresholds with SUMO motifs presented in LUAD..

We have generated plots from the prediction analysis to have a broader perspective on the genes mostly mutated to have SUMOylation sites. Below we have 2 figures. First figure shows the 10 genes that are the most frequently mutated genes to generate possible SUMOylation sites. The below figure consists of the length normalized genes, so the length of the protein is not influencing the data. The genes have changed completely when the parameter of length is removed.

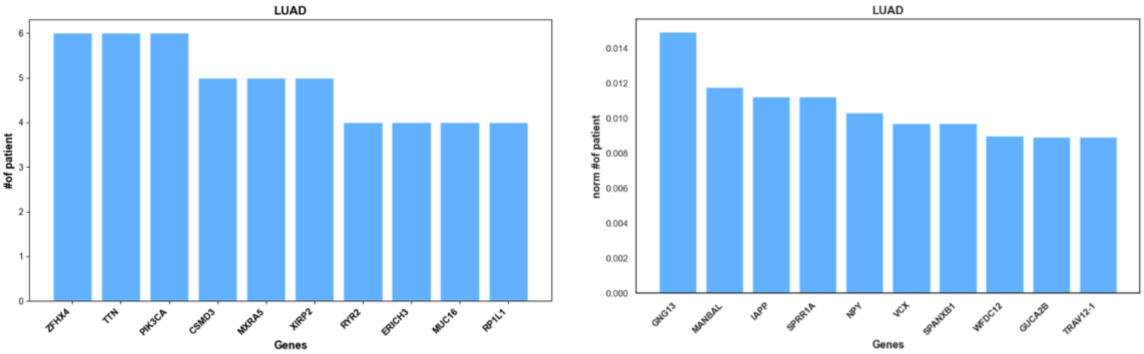


Figure 5. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length is considered for normalization in LUAD.

4.2 LUSC

We have also done prediction analysis via SUMOnet on mutation data of different cancer types. Aside from LUAD, we have analyzed the mutation data from LUSC cancer patients. Lung squamous cell carcinoma is another type of non-small cell lung carcinoma. GDC had 480 patient data and after processing the patient mutation data we have acquired 6246 subsequence data that had the mutation to lysine. From the prediction, in the 0.5 possibility threshold, 1035 of the subsequence positions were predicted to be a SUMOylation site. The table below summarizes the findings based on different thresholds. 30% of the positive predictions were also including SUMO motifs within the subsequence.

Threshold	# of predicted SUMO sites	# of including SUMO motifs
0.5	1035	313
0.7	526	204
0.9	204	105

Table 6. Possible SUMOylation sites predictions in different thresholds with SUMO motifs presented in LUSC.

The plots generated from the results of the predictions are given below. First plot is given the genes that are mutated the most and predicted to have a SUMOylation site. After normalizing the plot based on gene length, we obtained the second plot. Both plots have the “*pik3ca*” gene. The relevance of somatic mutations in *pik3ca* gene with cancer development is searched in other cancer studies. Based on its role in the cellular signaling pathways, the mutation can cause the cell growth and behavior to be uncontrolled and aggressive. This gene mutation can also be seen in many other cancer types._(Samuels et al., 2010)

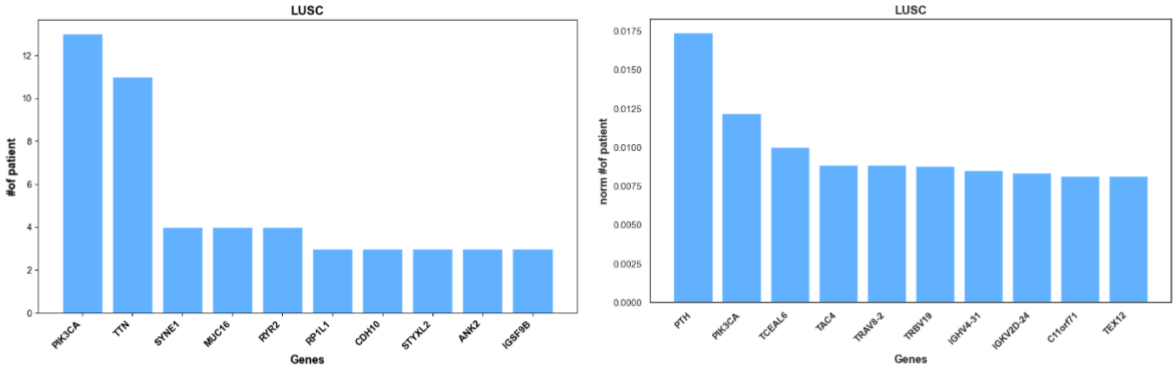


Figure 7. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length is considered for normalization in LUSC.

4.3 BRCA

We have gathered TCGA-BRCA data, representing breast cancer patient projects. We have run the pipeline on 778 breast cancer patients, moreover, we had 5376 mutated subsequences to get prediction from. From these 5376 peptide subsequences, SUMOnet predicted 999 to be a possible SUMOylation site with the threshold of 0.5. As we analyze the results in higher probability thresholds, at most 44% included a SUMO motif in the 0.9 threshold.

Threshold	# of predicted SUMO sites	# of including SUMO motifs
0.5	999	312
0.7	541	216
0.9	260	116

Table 8. Possible SUMOylation sites predictions in different tresholds with SUMO motifs presented in BRCA.

The gene analysis of our breast cancer studies gave an interesting result. Below we plotted the most frequently mutated genes that also was predicted to result in a SUMOylation site emergence. In both length normalized and non-normalized plots, "*pik3ca*" gene is the most prominent one.

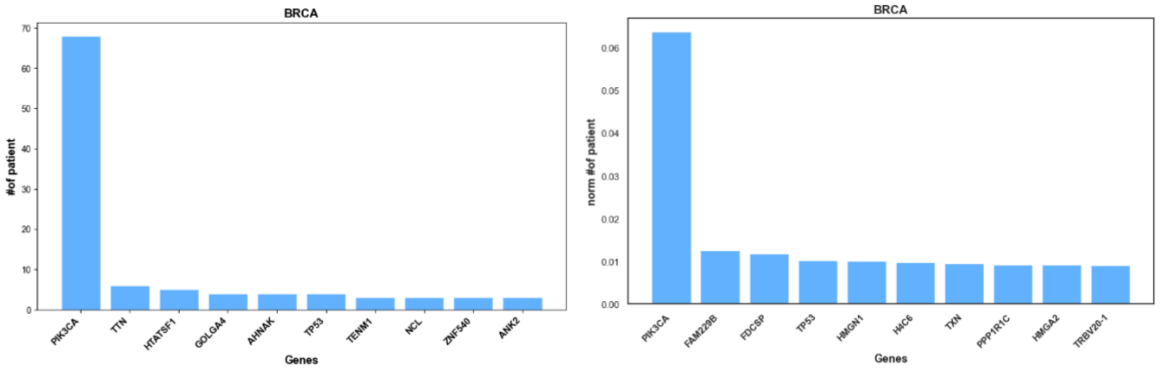


Figure 9. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length is considered for normalization in BRCA.

4.4 LGG and GBM

We have checked 2 types of cancer that have brains as their primary site. First one is the "Brain lower grade Glioma (LGG)" where the tumor cells arise from the glial cells, this is a slow progressing cancer type and 20% of the brain tumor are composed of this type. The second brain cancer is the "Glioblastoma Multiforme (GBM)". This is a very aggressive, malignant form of brain cancer and the spinal cord is in the affected areas.

GBM project data had 299 patients and the LGG project had 274 patients. We had 1669 subsequences where we had operated the SUMOnet on GBM data. The positive predictions to be a SUMOylation site was not very high. The motifs were present in 30% of the possible SUMO sites in 0.5 threshold. For the case of LGG, the patient number was 274, not very distant from GBM. However, the peptide subsequences we had acquired and processed to get ready for SUMOnet were very low with 656. The percentage of possible SUMO sites was higher than with GBM. On the other hand, motifs including subsequences were again 30% in the 0.5 threshold.

GBM

Threshold	# of predicted SUMO sites	# of including SUMO motifs
0.5	280	86
0.7	145	60
0.9	54	30

Table 10. Possible SUMOylation sites predictions in different tresholds with SUMO motifs presented in GBM.

LGG

Threshold	# of predicted SUMO sites	# of including SUMO motifs
0.5	114	34
0.7	66	27
0.9	29	14

Table 11. Possible SUMOylation sites predictions in different tresholds with SUMO motifs presented in LGG

Below plots are presenting the most frequently mutated genes with the results of possible SUMOylation sites. After normalizing the length of genes, in GBM's case only the "*arhgib*" gene is the common most mutated gene. For the LGG’s case, the non-normalized and length normalized plot gives “*tnfrsf13b*” to be the common one in both plots.

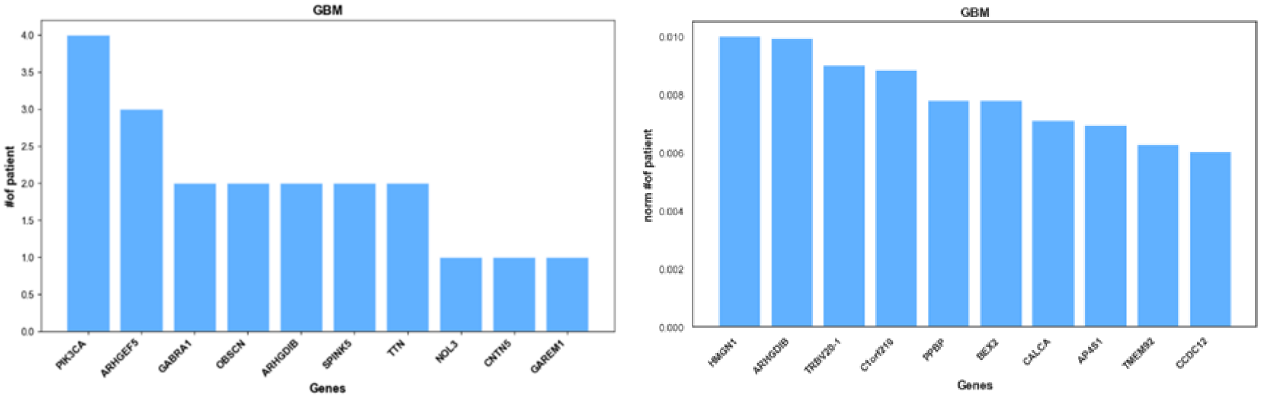


Figure 12. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length is considered for normalization in GBM.

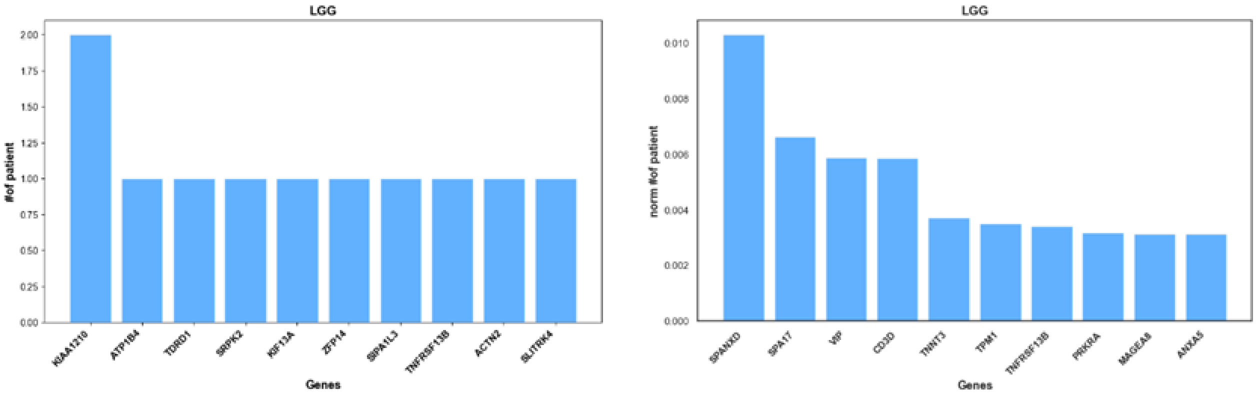


Figure 13. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length is considered for normalization in LGG.

4.5 STAD

Another cancer type we have analyzed is the stomach adenocarcinoma. The starting point was 400 patient mutation data, after preparing the subsequences from the given data we have acquired 3320 subsequences with Lysine mutation. After prediction via SUMOnet, 568 cases were predicted to be possible SUMO sites with 0.5 probability or higher. Among those 191 subsequences were containing SUMO motifs. As we narrow down the threshold the percentage of motifs containing subsequences increased.

Threshold	# of predicted SUMO sites	# of including SUMO motifs
0.5	568	191
0.7	286	129
0.9	118	70

Table 14. Possible SUMOylation sites predictions in different thresholds with SUMO motifs presented in each case in STAD.

From the obtained prediction results the most mutated genes with SUMOylation sites were gathered to generate the plots below. After normalizing based on the gene length, we have generated the second plot. Both plots have “*pik3ca*”, “*kcnmb2*” genes as most frequently mutated genes. *pik3ca* gene was mentioned before as being identified in many cancer types to be associated. The *kcnmb2* gene is studied in gastric/stomach cancer and was found to be downregulated in tumor cells, however, there is not enough evidence to link the two occurences. (Qiu et al., 2020)

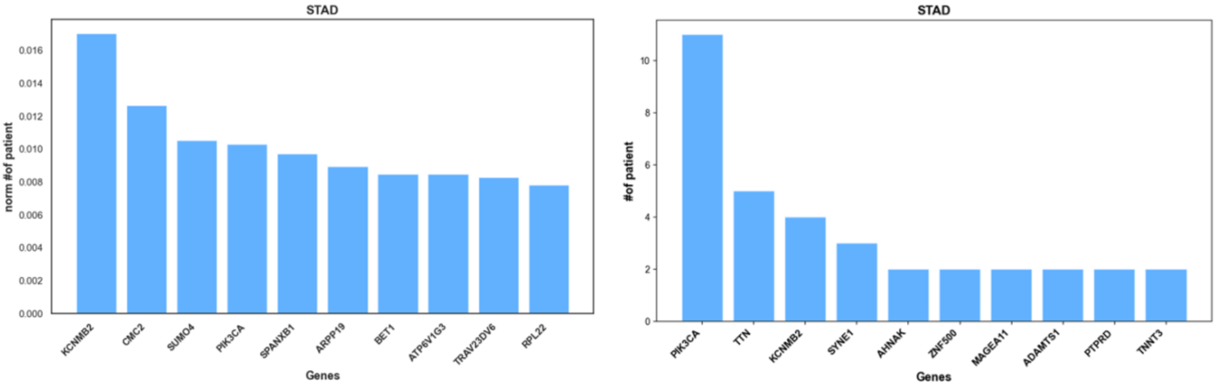


Figure 16. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in STAD.

4.6 Pan-cancer analysis

The plot below is generated from all the most frequently mutated genes that are also predicted to have SUMOylation sites data of different cancer types explained before. Most prominent gene is “*pik3ca*” and it is found in 3 of the 6 cancers.

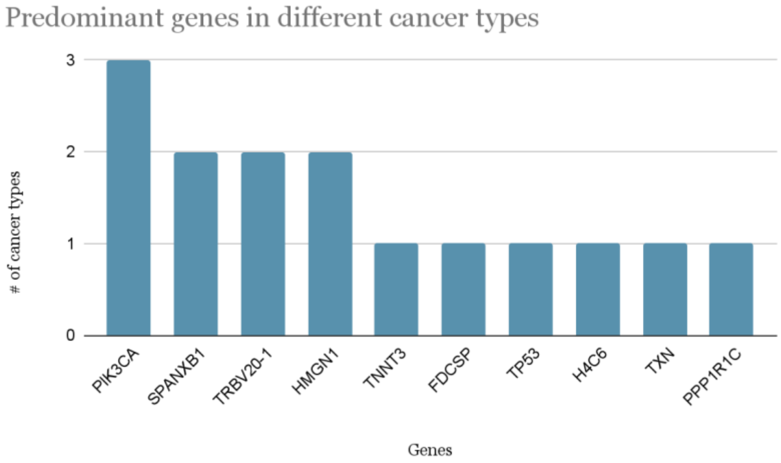


Figure 17. Most recurring genes that normalized in 7 cancer types discussed above.

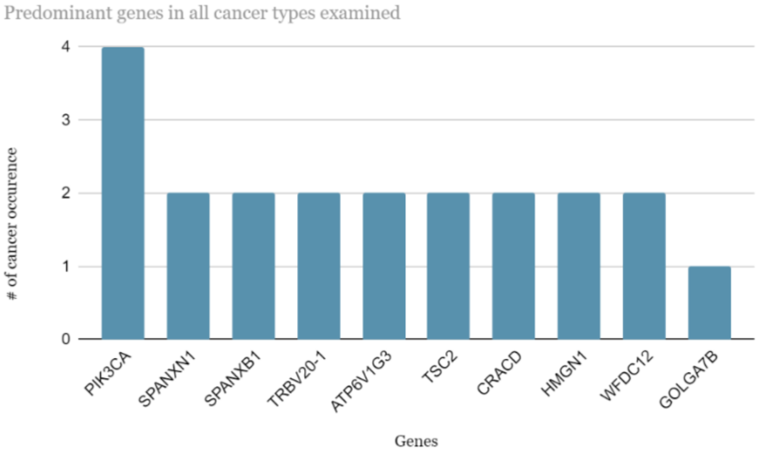


Figure 18. Most recurring genes that normalized in all 16 cancer types analyzed in the project.

4.7 Results of other cancer types

ACC

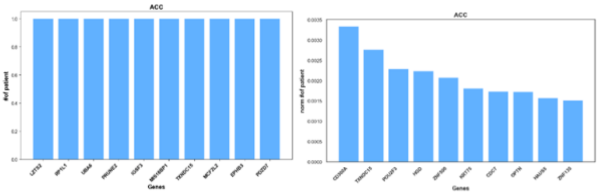


Figure 18. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in ACC.

KICH

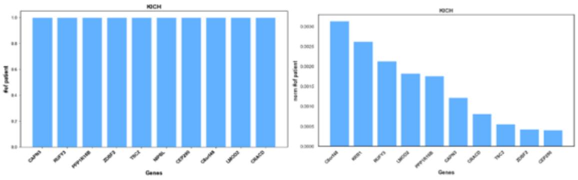


Figure 19. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in KICH.

MESO

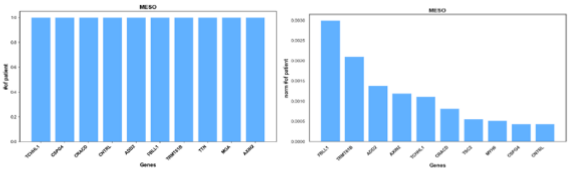


Figure 20. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in MESO.

OV

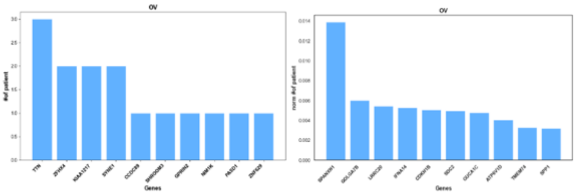


Figure 21. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in OV.

READ

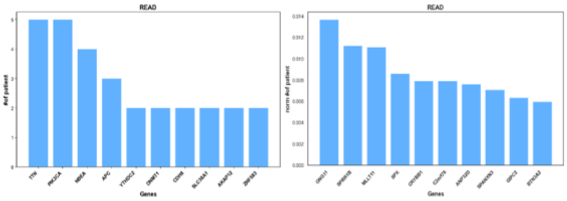


Figure 22. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in READ.

SARC

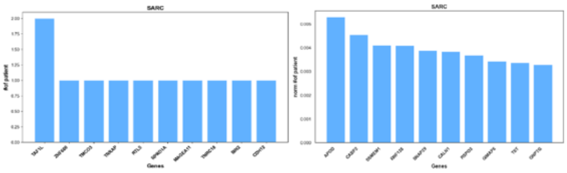


Figure 23. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in SARC.

SKCM

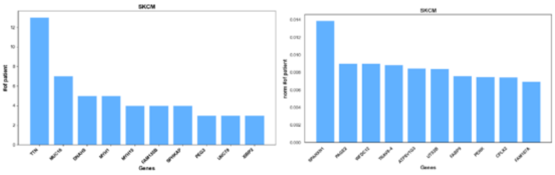


Figure 24. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in SKCM.

THYM

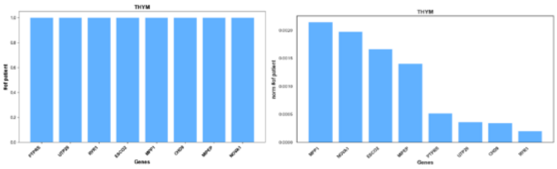


Figure 25. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in THYM.

UCS

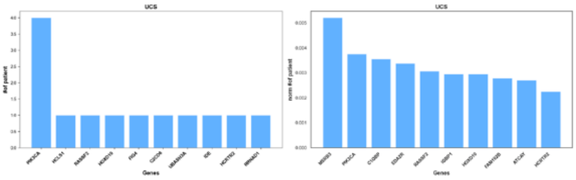


Figure 26. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in UCS.

UVM

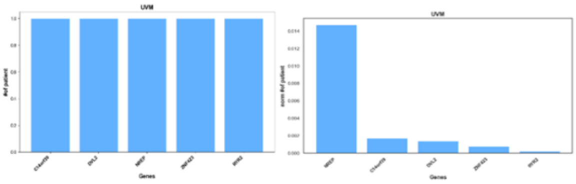


Figure 27. a) The most frequently mutated genes with likely SUMO sites, b) The most frequently mutated genes with likely SUMO sites when gene length normalized in UVM.

5 Conclusion

In this project, we aimed to develop a computational pipeline to detect new emerging possible SUMOylation sites due to somatic cancer mutations. We run our deep neural network SUMOnet on various cancer types and found several candidate mutations that are likely to be SUMOylated. The analysis of the most frequently mutated genes with predicted SUMOylation sites gave us few genes that are recurring on different cancer types. Moreover, focusing on those common genes and determining whether the mutation occurred in a similar position can help us understand and support the role of SUMOylation in cancer development. Future work will focus on a pan-cancer analysis and other corroborating evidence on expression and pathway enrichment that might lead to dysregulated and aberrant SUMOylation mechanisms. Also, the tlmsa main functions of the methods can be summarized as follows, the tlmsa function that provides mutation mapping mentioned in the section 3.3 is getMutatedseq, which can be easily used to obtain a mutated sequence and the corresponding functions of the tlmsa class for the generating subsequences and exploring the motifs that described in the section 3.4 is getSubseq and get_motif, respectively. All of the codes and functions described above has been published as an open-source project on GitHub in the scope of the package for usage and development by other researchers in their respective disciplines.

6 References

Wilkinson, K. A., & Henley, J. M. (2010). Mechanisms, regulation and consequences of protein SUMOylation. *The Biochemical journal*, *428*(2), 133–145. https://doi.org/10.1042/BJ20100158

Seeler, JS., Dejean, A. SUMO and the robustness of cancer. *Nat Rev Cancer* 17, 184–197 (2017). https://doi.org/10.1038/nrc.2016.143

Walsh C (2006) *Posttranslational modification of proteins: Expanding nature's inventory*. Englewood (CO): Roberts and Co. Publishers. xxi, p 490.

Beauclair, Bridier-Nahmias, Zagury, Saïb, Zamborlini, JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMs, *Bioinformatics*, Volume 31, Issue 21, 1 November 2015, Pages 3483–3491, https://doi.org/10.1093/bioinformatics/btv403

Samuels, Y., & Waldman, T. (2010). Oncogenic mutations of PIK3CA in human cancers. *Current topics in microbiology and immunology*, *347*, 21–41. https://doi.org/10.1007/82_2010_68

Qiu, J., Sun, M., Wang, Y., & Chen, B. (2020). Identification of Hub Genes and Pathways in Gastric Adenocarcinoma Based on Bioinformatics Analysis. *Medical science monitor : international medical journal of experimental and clinical research*, *26*, e920261. https://doi.org/10.12659/MSM.920261

Berke Dilekoğlu, Oznur Tastan. SUMONET: Deep Sequential Prediction of Sumoylation Sites, in preparation.

Antonio Colaprico, Tiago C. Silva, Catharina Olsen, TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data, Nucleic Acids Research, Volume 44, Issue 8, 5 May 2016, Page e71,c

Cokelaer et al. BioServices: a common Python package to access biological Web Services programmatically Bioinformatics (2013) 29 (24): 3241-3242