Detection of SUMOylation sites that emerge through Mutations in Cancer

STUDENTS / UNIVERSITIES

Suleyman Onur Dogan[1], Ecem Erdogan[2]

1. Department of Computer Engineering, Antalya Bilim University
2. Department of Molecular Biology and Genetics, METU

SUPERVISOR(S)

Oznur Tastan

INTRODUCTION

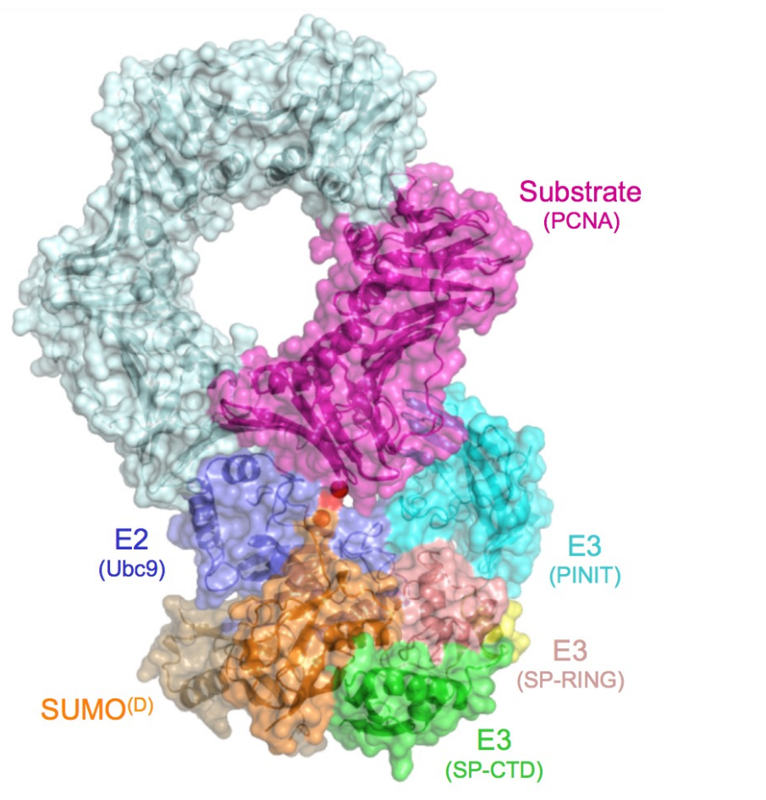

Figure 1. Regulatory proteins and SUMO protein of SUMOylation cycle

- **Post-translational modifications (PTMs)** are the addition or cleavage of subunits to proteins for the regulation, activation and localization. [1]
- **SUMOylation** is a PTM where Small Ubiquitin-like Modifier (SUMO) proteins are attached covalently and reversibly to target proteins at Lysine (K) residues. [2]
- SUMOylation regulation plays a key role in the nature of the cells.
- In cancer patients tumors undergo various somatic mutations.
- Some of these mutations may alter the protein sequences and novel SUMOylation sites may emerge.

PROBLEM STATEMENT

Our aim is to develop a computational pipeline to detect emerging likely SUMOylated sites.

Wild type sequence: LVSHVMACGGSTLYPGFTKRL

↓ ?

Cancer patient sequence: LVSHRMACGGKTLYPGFTKRL

SUMOylated or not

METHODS

Mutation Data from GDC:

- Mutation data of LUAD patients retrieved from GDC via TCGA packages. [3]
- Since, SUMOylation occurs on lysine residues, mutation data is filtered by finding patients' genes that has mutation resulted in K.
- Also, mutations nearby mutated lysine may affect the SUMOylation. Therefore, we considered all mutations of the corresponding patients/genes.

Sequence from UniProt:

- Protein sequences are obtained from UniProt by using Python Bioservices package [4].

Mapping Mutation to sequence:

- Patients' mutations mapped to wild type peptide sequence; thus, mutated peptide sequences have been obtained for each patients' protein.

SUMOylation prediction:

- We used SUMOnet to predict SUMOylation sites. SUMOnet is a deep neural network to predict possible SUMOylation sites from given 21 long protein subsequence with target K in the middle. SUMOnet achieves 87% AUC score [5].

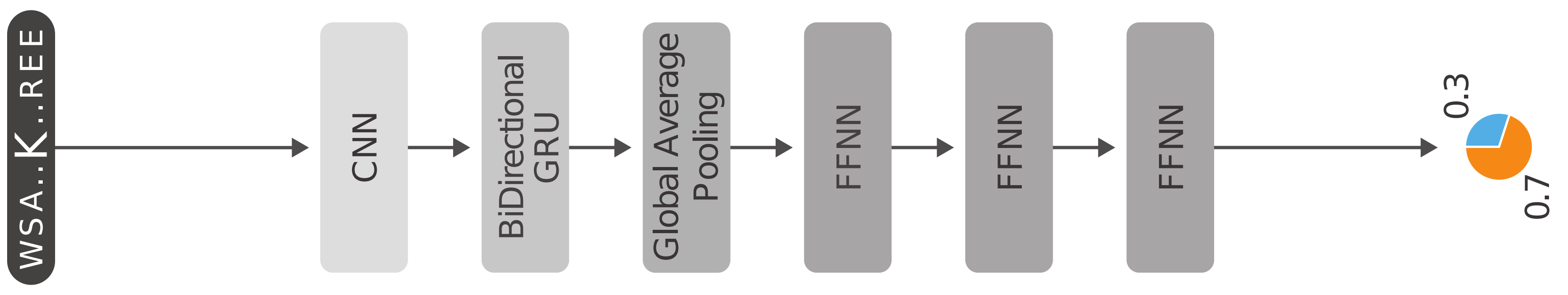

Figure 2. Architecture of deep neural network of SUMOnet

METHODS

Subsequence for SUMOnet:

- Since, input shape of SUMOnet is 21 long subsequence, mutated peptide sequence is created by finding mutated lysines and getting subsequence nearby mutated lysines.
- In special cases where there is no 10 amino acid before or 10 amino acid after mutated lysines, empty parts has been filled with X.

For each patient's gene:

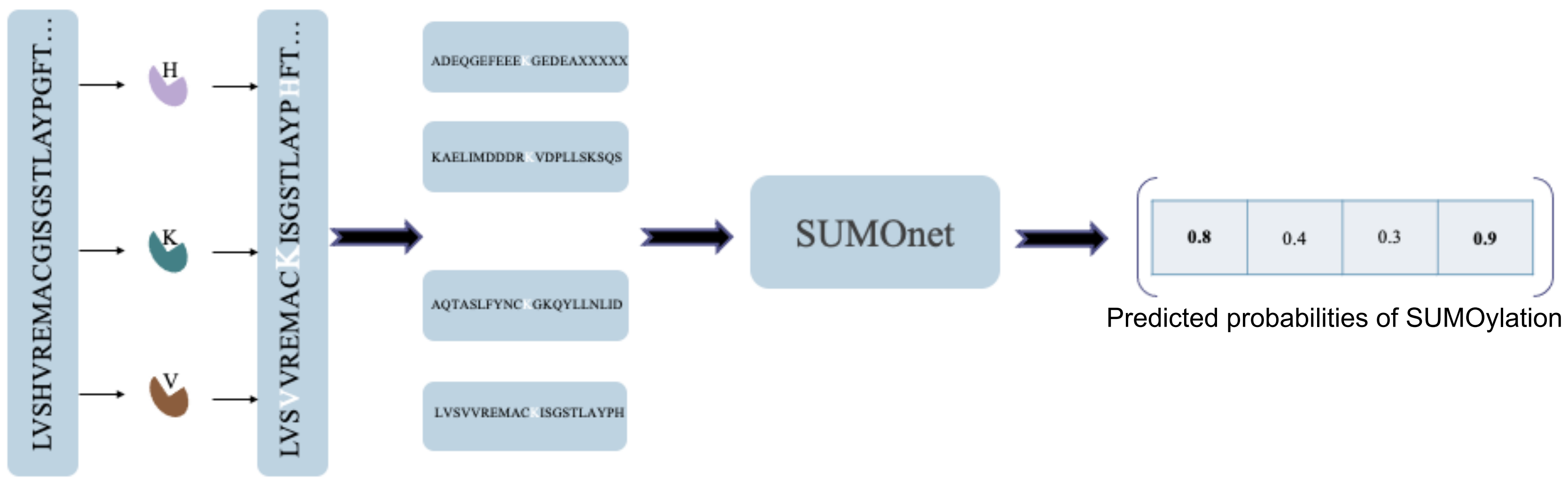

Figure 3. Summarized workflow

RESULTS

- We ran the pipeline on 521 LUAD patients.
- There were 6661 candidate peptide sequences centered on mutated K.
- Below we report the possible number of SUMOsites for various threshold for SUMOylation probability. We also report how many of these candidate sites resides in known SUMOylated motifs.[6]

Cutoff	# of predicted SUMOsites	# of including SUMO motifs
0.5	1059	349
0.7	514	230
0.9	191	111

Figure 4. Possible SUMOylation sites predictions in different tresholds with SUMO motifs presented in each case.

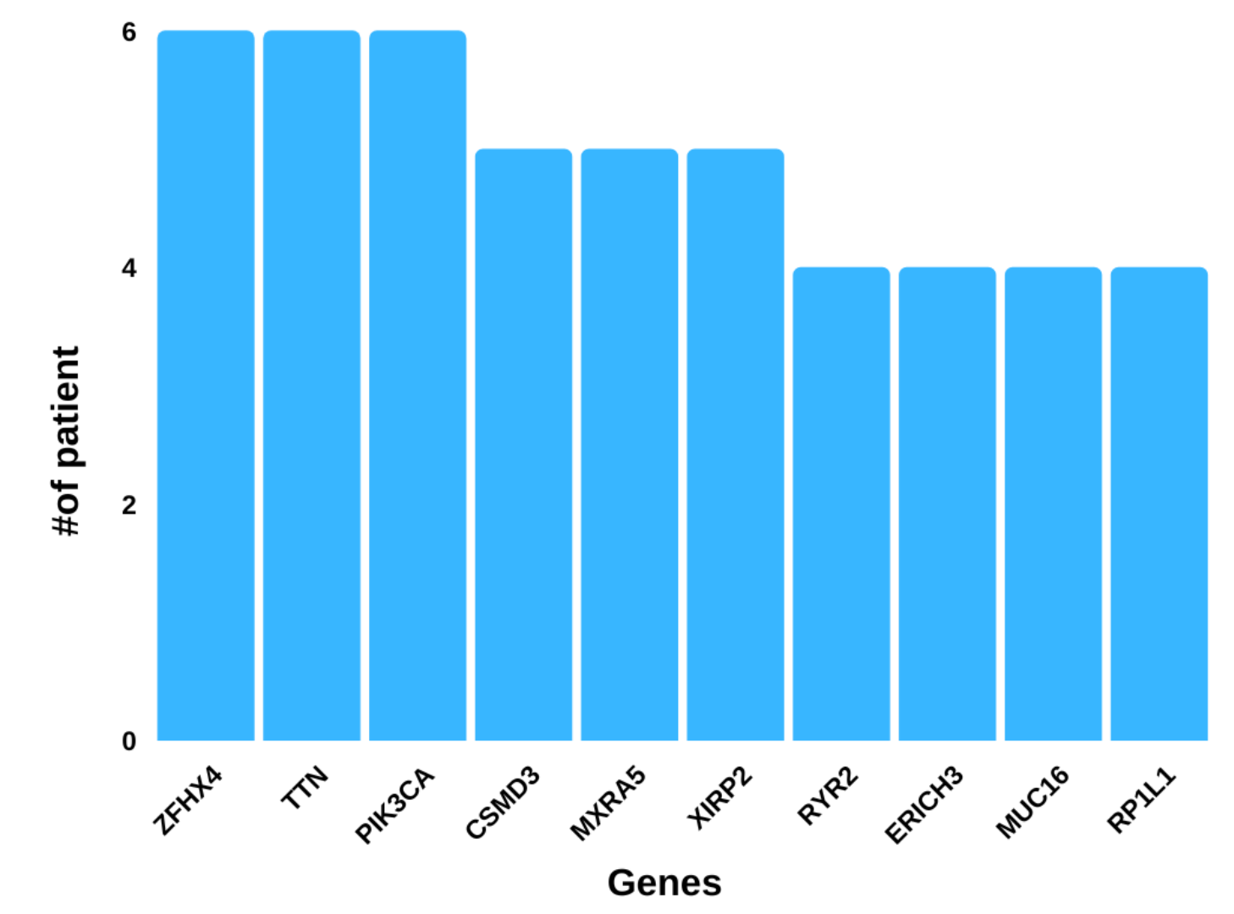

Figure 5a. The most frequently mutated genes with likely SUMO sites.

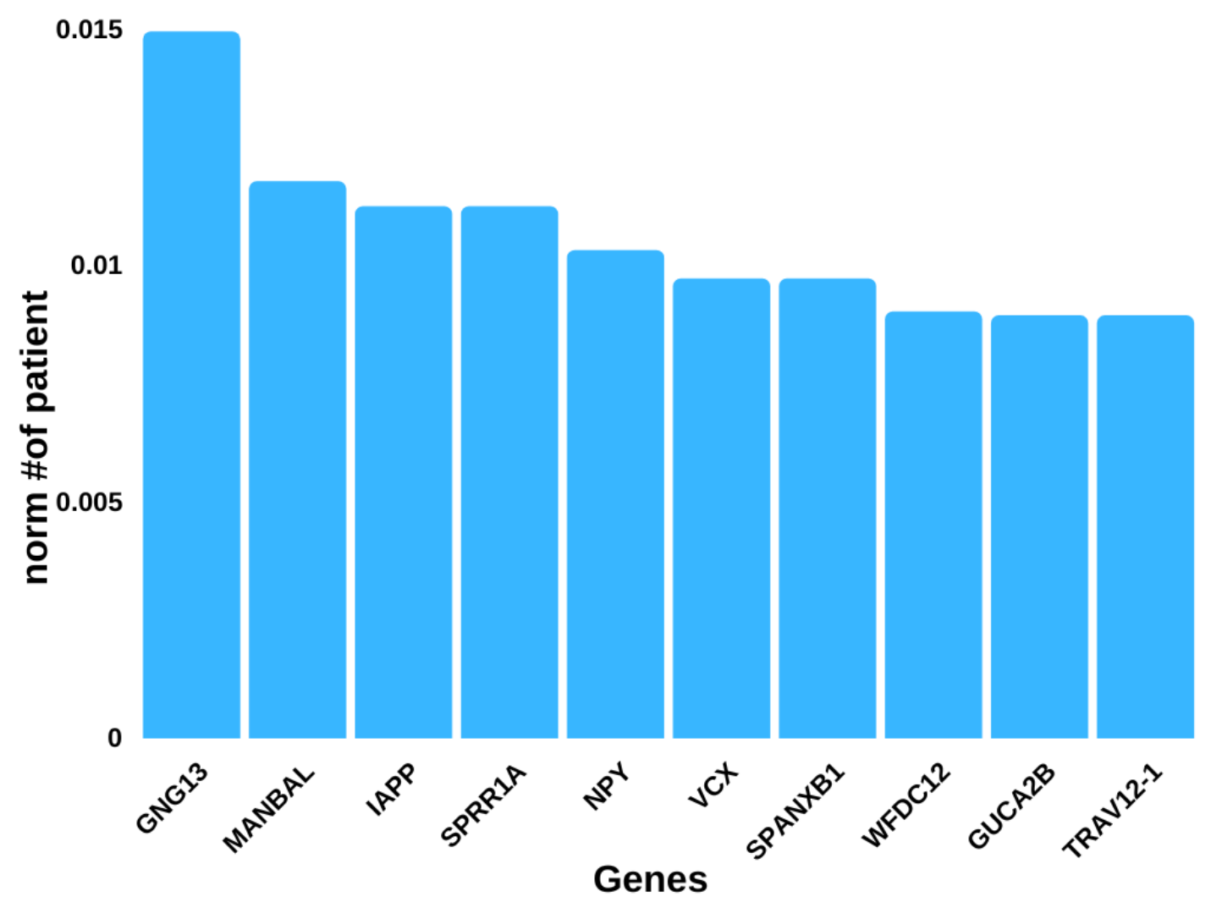

Figure 5b. The most frequently mutated genes length normalized with SUMO sites.

CONCLUSIONS

- We aimed to reveal the probability of cancer progression with mutations creating new SUMOYlation sites, thus a dysregulated and aberrant SUMOylation mechanism is operating.
- LUAD cancer patient data was retrieved from GDC portal, then the positions of the mutations was mapped to the non-mutated whole sequence retrieved from UniProt.
- Subsequence was generated as 21 amino acid long and Lysine residue residing in the middle
- For the purpose of SUMOylation prediction, we employed SUMOnet.
- All the codes are published as an open-source project in GitHub with functions.

github.com/sonurdogan/tlmsa

REFERENCES

[1] *Overview of post-translational modification: Thermo Fisher Scientific - US*. Overview of Post-Translational Modification | Thermo Fisher Scientific - US. (n.d.). Retrieved August 12, 2022, from https://www.thermofisher.com/tr/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-post-translational-modification.html

[2] Geiss-Friedlander, R., & Melchior, F. (2007). *Concepts in sumoylation: A decade on*. Nature News. Retrieved August from https://www.nature.com/articles/nrm2293

[3] 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

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

[5] Berke Dilekoğlu, SUMONET: Deep Sequantial Prediction of Sumolaytion Sites, Master's thesis

[6] Beauclair, G., Bridier-Nahmias, A., Zagury, J.-F., Saïb, A., & Zamborlini, A. (2015, July 2). *Jassa: A comprehensive tool for prediction of Sumoylation sites and Sims*. OUP Academic. Retrieved August 29, 2022, from https://academic.oup.com/bioinformatics/article/31/21/3483/195061