Biological Datasets for Computational Physics

Cancer mutations in Sororin protein - Q96FF9

Matteo Pedrazzi - 2076719

(Dated: June 25, 2023)

The Sororin protein, encoded by the CDCA5 gene, plays a critical role in maintaining genomic stability and regulating cell division. Alterations in Sororin due to cancer-associated mutations have been implicated in various types of cancers, but a comprehensive analysis of these mutations is still lacking. The goal of this project is to conduct a detailed investigation of cancer mutations in the human Sororin gene CDCA5, through the analysis of publicly available cancer genomics databases, identifying the spectrum of cancer mutations in CDCA5 across the full sequence and trying to study how these mutations are involved in tissue-specific effects and their association with distinct cancer types. The distribution of these mutations is also studied with respect to structural properties of the sequence and the predicted secondary structures, revealing recurrent mutations and potential hotspot regions within the protein sequence, shedding light on critical residues susceptible to cancer-associated alterations, trying to help in the search of potential links that would contributes to a better understanding of the molecular mechanisms underlying tumorigenesis.

INTRODUCTION

Sororin is a protein encoded in the CDCA5 (Cell division cycle-associated protein 5) gene in humans (Q96FF9, also known as p35), but present also in other eukariote organism, like mice. It is a relatively small protein, consisting of 252 amino acids and the name Sororin, after the Latin word 'soror', which means 'sister', derives from its initial discovery in Drosophila melanogaster, where it was observed to be essential for sister chromatid[1] cohesion in mitosis, stabilizing cohesin complex association with chromatin [1–3], but how Sororin performs these functions is still unknown even if it is supposed to be conserved in different species, indicating its fundamental importance in cellular processes.
During mitosis and meiosis, when cells divide to produce two identical daughter cells, the proper separation of sister chromatids is critical to maintain genomic stability. Sororin plays a vital role in this process by ensuring the cohesion between sister chromatids until the appropriate time for their separation, preventing premature separation and premature loss of chromosomal integrity.
Sororin's role in ensuring proper chromosome segregation and maintaining genomic stability suggests its potential involvement in tumor growth and metastasis in cancers: the dysfunction or depletion of Sororin can lead to severe defects in chromosome segregation, resulting in aneuploidy (abnormal number of chromosomes) and genomic instability, which can fuel the development of genetic diversity within tumors, leading to increased tumor heterogeneity and potentially facilitating metastasis, the spread of cancer cells to other organs.

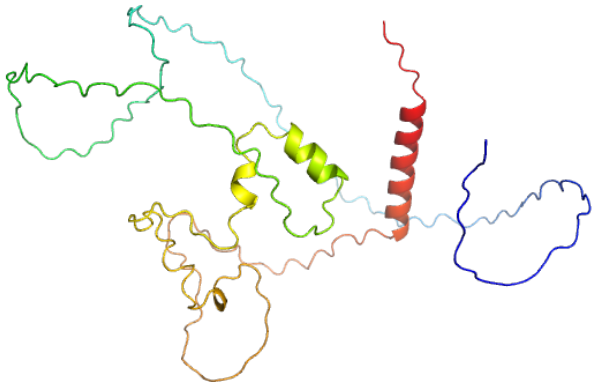


FIG. 1: *Prediction for the 3D structure of Sororin by AlphaFold [9, 10], visualized using PyMOL [11].*

Genes are contained in chromosomes, which are long strands of DNA in each cell, and they are responsible for the coding instructions of proteins. Each chromosome has many different genes. CDCA5 is a gene on the chromosome 11[2], and according to the related literature [1, 3] immunofluorescence analysis revealed the subcellular location of Sororin is primarily the nucleus, observing that it associates with nuclear chromatin from S phase until metaphase and is released in the cytoplasm upon nuclear envelope breakdown[2] explaining diffuse distribution of CDCA5 throughout the cell during mitosis.

[1] sister chromatids are the identical copies formed by the DNA during the replication of a chromosome.

[2] Hartz (2011) mapped the CDCA5 gene to chromosome 11q13.1 based on an alignment of the CDCA5 sequence (GenBank BC011000) with the genomic sequence (GRCh37).

Any change in the DNA sequence of a cell is a mutation and it may be caused by mistakes during cell division or by exposure to DNA-damaging agents in the environment. Mutations can be beneficial or have no effect, but they can also be harmful, leading to cancer or other diseases. The purpose of this project is to study the mutation landscape for the Sororin protein and see how this is related to the diseases associated to CDCA5, starting from the mutations and cancers visualization along the entire sequence and finishing with some structural insights, like the secondary structure and the region or motifs of the sequence. These subregions, detected both from structural and functional properties, are taken into account to have a more complete description of the sequence and thus to be able to better understand mutations in each area.
Looking at genetic disorders and diseases can contribute to understanding the underlying mechanisms and developing targeted treatments. This paper will simply proceed though the enunciated arguments, starting on one side from the clinical point of view with the frequency of the different cancer types, on the other hand we will try to converge to the same conclusions starting from available online data, looking at single-point mutations and their associated features.

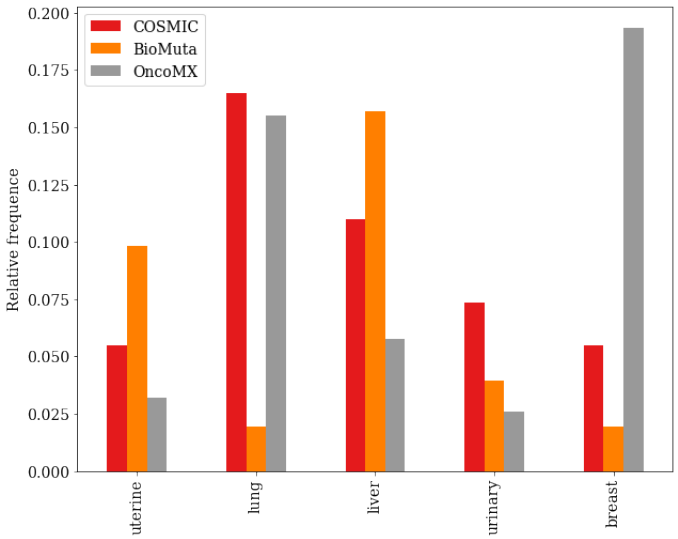


FIG. 2: *Histograms reporting the number of tumorous samples originated from Sororin mutations in different datasets, normalized over the total number of samples in each dataset, in order to obtain comparable values.*

METHODS

Datasets

The 252 amino acids sequence composing the primary structure of Sororin CDCA5 human protein is retrieved in FASTA format from the UniProtKB database [4]. The other sources, used in different type of analysis and regarding the mutations with their associated diseases, are all publicly available online datasets:

- Catalogue Of Somatic Mutations In Cancer (COSMIC v98) [5]: containing different datasets, one of them has a complete overview of the present mutations (position, nucleotide change, amino acid change, frequency and type) and will be used for Fig. 3, the others include also the affected tissue (primary and eventually also secondary) and other useful features;

- DisGeNET [6]: summary of observed Gene-Disease Associations (GDAs), assessing for type and class for each of the available diseases, as well as other columns describing the number of genes associated to the disease, the score of the association and the year of first and last reference in literature;

- BioMuta and BioXpress [7]: the first contains a summary of single-nucleotide variations (SNVs) and gene expression in cancers, with a total of 38 rows with 37 unique variations. The second one instead relates the different cancer types to their subject ratio and other statistical features, including the expression trend. However the BioXpress dataset used is the one accessible from the OncoMX [8] portal (under the entry BioXpress), which is more complete and easy to use.

Analysis Methods

Other online tools have been exploited for the analysis of some of interest physical-chemical properties. Some properties can be directly inferred from the primary structure, using the Expasy ProtScale [12] website, which allow to compute quantities of interest for the whole amino acid sequence, based on different techniques. The measures taken in into account to have an insight on the structural features of this sequence are hydropathicity [13], polarity [14] and average flexibility [15], considering the algorithm running for a window size of 7 amino acids and taking the normalized scores in $[0, 1]$. The probabilistic prediction of the secondary structure have been done using GOR IV prediction method [16, 17].

Statistical analysis

DisGenet also collects some meaningful indexes and measures, like Disease Specificity index (DSI) and Disease Pleiotropy index (DPI), or the probability of being

loss-of-function intolerant (pLI)[18]. The first two are defined as follows [19]:

$$DSI = \frac{\log_2(N_d/N_T)}{\log_2(1/N_T)} \quad (1)$$

$$DPI = \frac{N_{dc}}{N_{TC}} \cdot 100 \quad (2)$$

where N_d and N_{dc} are the number of diseases and different MeSH disease classes of the diseases associated to the gene/variant, while N_T and N_{TC} are the total number of diseases and MeSH diseases classes in DisGeNET.
Python code have been developed for datasets inspection and statistical analysis, all the plots are done using the Matplotlib library[3].

RESULTS

As anticipated in the methods section we will start from the cancer frequency in the inspected datasets, then we will go through all the presented datasets to see which are the hotspots for mutations and their possible linkages with cancers. Finally also functional features and the secondary structure are analysed.

Cancer affected tissues

Sororin can have mutations leading to different types of diseases, here we want to put the attention on the cancer diseases. The distribution of the different developed cancer types in the Sororin protein, could be useful to see which are the dominant cancer diseases emerging from the patients tested on the available datasets. In particular from Fig. 2 we have a comparison of the cancers in the 3 datasets used for this purpose. We have that cancer types with a higher incidence, resulting from Sororin mutations, are different for each of them and we can clearly observe Cosmic, BioMuta and OncoMX are quite unbalanced with respect to each other. This suggests different datasets can be optimal if one is interested in a certain tumour disease and in that case it would have been useful to better exploit the datasets content, but in general the results are not comparable at all. The same can be said for the features described in each of them: in the following we'll use specific datasets for specific task in order to exploit their peculiarity.
Together with complete plots for Cosmic, BioMuta and OncoMX, a different plot is reported in the appendix for DisGeNET, considering instead the number of genes leading to certain cancer type. From DisGeNET are also provided values for DSI equal to 0.563, DPI equal to 0.808, meaning the number of diseases associated to this gene is not huge and they represent a good portion of the existent MeSH diseases classes, while pLI around 0.0011 means the gene is extremely tolerant with respect to loss of function variations.

Mutation landscape

In the wide field of mutations, acquired gene mutations are a much more common cause of cancer than inherited mutations. Among the acquired ones, our attention

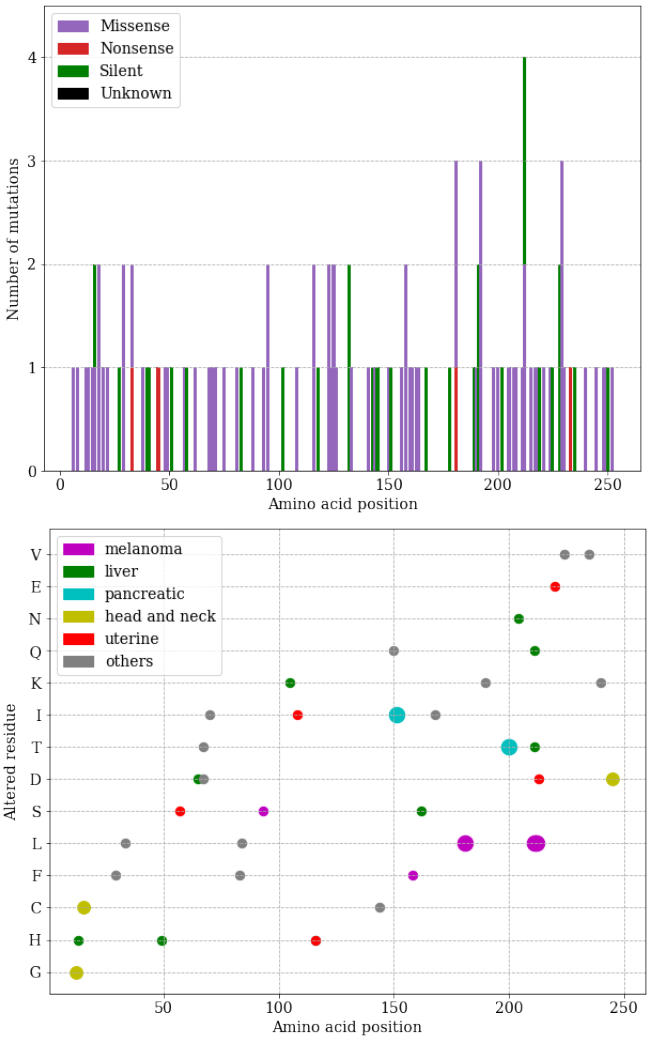


FIG. 3: *Visual representation of cancer mutation in COSMIC (top) with frequencies vertical lines mapped onto the amino acid sequence and representation for the mutated amino acids vs. position on the sequence for BioMuta (bottom) having dots sizes proportional to the frequency of the mutations and colored using the 5 most frequent cancers in this dataset.*

[3] The code used is available at the following GitHub public repository: https://github.com/matteopedrazzi/BDCP_Project.

will focus on the single-point mutations (single-nucleotide variations, SNVs), leaving apart possible insertion or deletion from the primary structure. Nucleotides mutations can lead to different encoding issues in the synthesization of proteins, based on the effect they have there exist three different kinds of point mutations [19]:

- silent: nucleotide change at DNA level has effect on the RNA replica but the synthesized protein is the correct one;
- nonsense: change in the sequence causes no protein synthesis;
- missense: wrong protein produced, can be conservative or either non-conservative.

We are now going go make some statistics on the type and on the frequency of nucleotides and amino acids variations present in the datasets. From this analysis we would aim to answer to the initial question, asking whether is possible or not to identify mutation responsible for some kind of cancer, searching for a linkage between the observed mutation and involved tissue. Using the different datasets presented in previous section we can get an idea of the distribution of all types of cancer mutations along the sequence and also of their frequency, identifying potential hotspot sites.
Table I shows most of the mutations are missense substitutions, followed by synonymous, unknown type and nonsense. Observing the top plot in Fig. 3 we can observe there is a trend concentrating mutations on the final part of the sequence, in particular the 212-th amino acid is the most frequently mutated, but if we look only at missense the 192 (2 different possible mutations) and the 229 have the higher number of recorded variations. From the lower plot in Fig. 3 one can see for example the 'head and neck' cancer type is caused by mutations on the edges of the sequence, while the other types displays do not show a clear trend. The same plot could be reproduced on a larger scale for all the different cancer types to have a complete overview.

Through a further analysis is also possible to compute some statistics regarding the types of the observed mutations, considering the single residual changes, in terms both of nucleotides and amino acids. In particular for the latter, it's not really informative to have a statistics for the amino acid changes from one to another, given that the amount of possible combinations is huge, but rather let's look at the mutations from one family/cluster of amino acids to another, which can tell us something more interesting about the mutations happening. Table II is showing quite surprisingly that the mutations '$C > A$', '$C > T$' and '$G > A$' enclose almost 75% of the total possible mutations. Instead Fig. 4 is a kind of family transition matrix that show frequency of residual changes from one family of amino acids to another. Outside the diagonal, describing the conservative mutations,

Mutation type	Number of samples (%)
Nonsense substitution	4 (2.78)
Missense substitution	74 (51.39)
Synonymous substitution	28 (19.44)
Other	21 (14.58)

TABLE I: *Overview of the types of mutation in the COSMIC database given a total of 144 unique samples.*

there are some interesting patterns, like the polar to non-polar (and viceversa) residual changes, and generally all the other residues seem to be likely to mutate into polar ones.

Structural and functional properties

Finally, an analysis can be conducted on the observed sequence variations with respect to the sequence itself, with its characteristic functional domains, but also with respect to the properties of the primary structure, like measures before mentioned in methods section and the secondary structure.
We can gain some insights about a protein function by identifying sequences of amino acids that are already known to have a specific function. Regions are widespread and conserved patterns of interest that play a role in protein function/structure. Another recurrent type are compositional biases, referring to the uneven distribution of amino acids or specific amino acid motifs within the protein sequence. Biases in amino acid composition might indicate functional or structural significance.

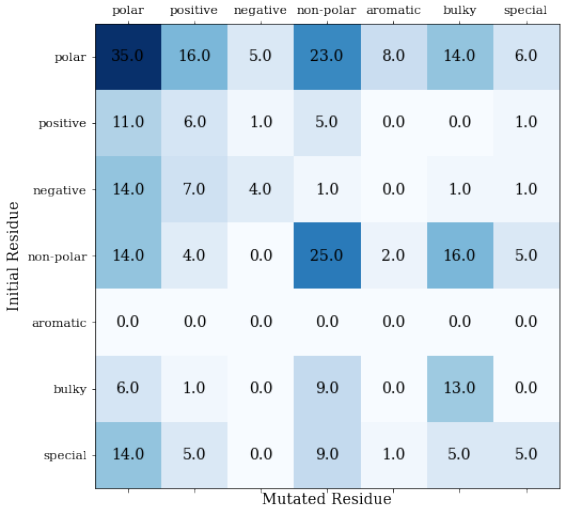


FIG. 4: *Family transition matrix, recording the statistics for all the mutations in terms of residues families. See Fig. 8 in the appendix for an overview of the initial amino acids distributions and their families.*

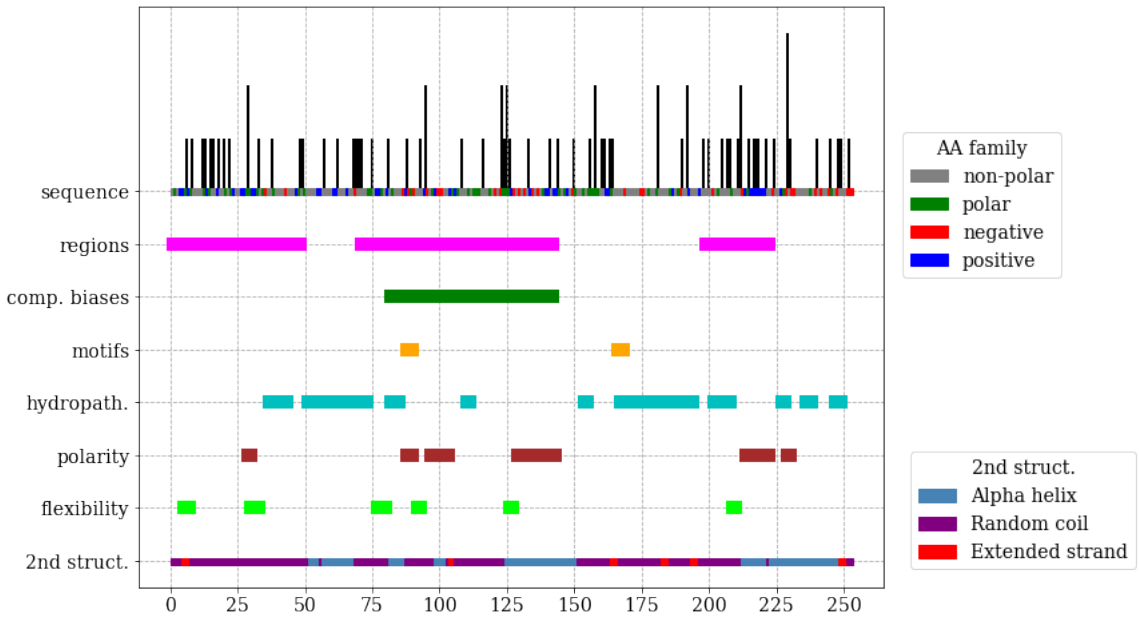


FIG. 5: *All the computed measures with overimposed mutations on top. On the right the legend related to the AA families in the sequence and the secondary structures.*

Mutation type	# samples (%)
A>C	1 (0.98)
A>G	3 (2.94)
A>T	4 (3.92)
C>A	14 (13.73)
C>T	36 (35.29)
C>G	5 (4.90)
G>A	26 (25.49)
G>C	4 (3.92)
G>T	8 (7.84)
T>C	2 (1.96)
T>G	1 (0.98)

TABLE II: *Overview of the nucleotides involved in the mutations observed in the COSMIC database given a total of 102 unique samples, considering there are 2 substitutions where the nucleotide change is unknown.*

Also short motifs are reported (usually not more than 20 amino acids), indicating specific binding or reactive sites of biological significance. As visible from UniProt, under the 'Features' section (Family & Domains, reported in Tab. III here), and in agreement with Expasy ProtScale results, are identified 3 regions, both of them disordered, 3 compositional biases which are almost overlapping and consisting of basic and acidic residues the external ones while polar residues in the middle and finally 2 short motifs, the KEN box (88-90) and the FGF motif (166-168). Also some properties can be derived from the primary structure are taken into account, namely hydropathicity, polarity and average flexibility as well as the predicted secondary structure using GOR IV method.

Fig. 5 is the comprehensive plot for all the feature above mentioned, with segments for the Protscale results obtained filtering the values where the score is above 0.5, and from these we can appreciate the higher values of hydropathicity and polarity are one the inverse to the other, which is reasonable because hydropathicity indicates the relative hydrophobicity or hydrophilicity of an amino acid residue. We can see flexible regions are located in regions that are likely to have more mutations, while observing the secondary structure we have that predicted Alpha helices are in correspondence with polar regions (low hydropathicity) and highly flexible regions are in correspondence with predicted coils. They are not fully visible from Fig. 5, but comparing some other figures one could also detect correlations between the kind of mutations (or the cancer type) with the properties of the regions in which they happen. An example could be the possible presence of mutations leading to uterine cancers in the regions with higher flexibility, describing the symmetric/asymmetric distribution of amino acid residues in the protein molecules.

DISCUSSION AND CONCLUSION

This project had the purpose to see if is it possible to establish a relation between mutations observed in Sororin CDCA5 protein and the cancer diseases associated. Lots of results and lots of possible other techniques

Type	AA Position	Description
Region	1-48	Disordered
Region	71-142	Disordered
Compositional bias	82-103	Basic and acidic residues
Motif	88-90	KEN box
Compositional bias	107-122	Polar residues
Compositional bias	125-142	Basic and acidic residues
Motif	166-168	FGF motif
Region	199-122	Disordered

TABLE III: *Regions, compositional biases and motifs highlighted by UniProt along the entire Sororin sequence, with their position and description.*

can be developed, here it's only a brief overview of the topic by a non-biologist, without the full comprehension of the underlying biological and associated genetic context.

While studies in literature are also focused on the importance of post translational modifications (PTM) in tumorigenesis, like Ser209 posphorilation [22], here the analysis is trying to answer different questions and maybe the conclusions can't be fully satisfying, because also in literature the binding partner of Sororin are well known, but the identification and characterization of the precise binding sites and the residues involved in these interactions are subjects of ongoing investigations.

Even if my analysis hasn't showed particular results regarding the KEN box, we can discover from the Eukariotic Linear Motif (ELM) [20] resource, that the KEN box is required for the association with the APC/C complex, in particular Sororin is targeted for APC/C-mediated degradation via the KEN box motif. Also the FGF motif revealed to alter interaction with PDS5, because both Sororin and WAPL proteins contain FGF motifs and function against each other by competing in binding to Pds5 via FGF motif [21]. In another work by *Zhang & Pati* [22] was found out the importance of the C-terminal Sororin domain, which is responsible for the interaction with cohesin in sister chromatid cohesion. Like we observed in the results section, the final part of the sequence contains several possible mutations, about which Zhang proved that mutating two of the last 16 amino acids into alanine (F241A, F247A) severely reduces the interaction of Sororin and cohesin, whereas deletion of the last 39 amino acids (from 214 to 252) of Sororin completely abrogates the Sororin-cohesin interaction.

CONCLUSION

Sororin CDCA5 could be an important potential prognostic indicator and target in certain kinds of cancers for which CDCA5 gene is overexpressed. This paper has its physiological space and time limitations, like the fact that only few measures have been considered, and also the limited number of mutations and cancer types deeply inspected. Applying the same concepts on a larger scale could lead to interesting results and development of new useful knowledge in this field, that can be used to tackle more specific problems with higher efficacy.

[1] Rankin S, Ayad NG, Kirschner MW. Sororin, a substrate of the anaphase-promoting complex, is required for sister chromatid cohesion in vertebrates. Mol Cell. 2005 Apr 15;18(2):185-200.

[2] Schmitz J, Watrin E, Lénárt P, Mechtler K, Peters JM. Sororin is required for stable binding of cohesin to chromatin and for sister chromatid cohesion in interphase. Curr Biol. 2007 Apr 3;17(7):630-6.

[3] Nishiyama T, Ladurner R, Schmitz J, Kreidl E, Schleiffer A, Bhaskara V, Bando M, Shirahige K, Hyman AA, Mechtler K, Peters JM. Sororin mediates sister chromatid cohesion by antagonizing Wapl. Cell. 2010 Nov 24;143(5):737-49.

[4] The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, Pages D523–D531.

[5] John G Tate and others, COSMIC: the Catalogue Of Somatic Mutations In Cancer, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D941–D947.

[6] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, Laura I Furlong. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucl. Acids Res. (2019)

[7] Dingerdissen HM, Torcivia-Rodriguez J, Hu Y, Chang TC, Mazumder R, Kahsay R. BioMuta and BioXpress: mutation and expression knowledgebases for cancer biomarker discovery. Nucleic Acids Research, gkx907. 2017 Oct 09.

[8] Dingerdissen HM, Bastian F, Vijay-Shanker K, Robinson-Rechavi M, Bell A, Gogate N, Gupta S, Holmes E, Kahsay R, Keeney J, Kincaid H, King CH, Liu D, Crichton DJ, Mazumder R. OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data. JCO Clin Cancer Inform. 2020 Mar;4:210-220.

[9] Jumper J, Evans R, Pritzel A et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

[10] Mihaly Varadi and others, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D439–D444.

[11] Schrödinger L & DeLano W, 2020. PyMOL, Available at: http://www.pymol.org/pymol.

[12] Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A; Protein Identification and Analysis Tools on the Expasy Server; (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005). pp. 571-607

[13] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982 May 5;157(1):105-32.

[14] Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974 Sep 6;185(4154):862-4.

[15] Bhaskaran R and Ponnuswamy PK (1988), Positional flexibilities of amino acid residues in globular proteins. International Journal of Peptide and Protein Research, 32: 241-255.

[16] Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol. 1996;266:540-53.

[17] Combet C, Blanchet C, Geourjon C, Deléage G. NPS@: network protein sequence analysis. Trends Biochem Sci. 2000 Mar;25(3):147-50.

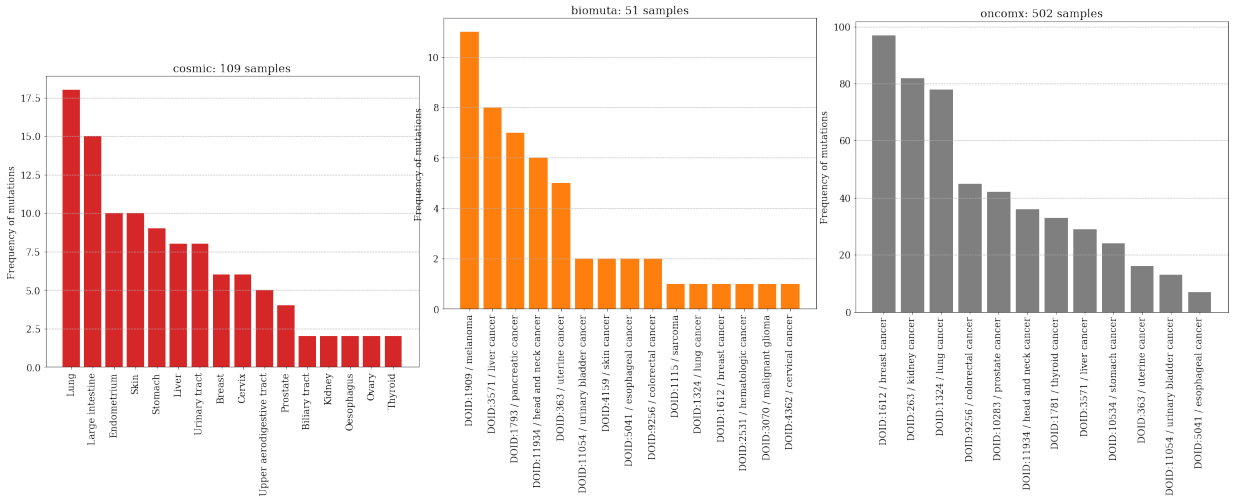
[18] Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. Nat Genet. 2019 May;51(5):772-776.

[19] Fuxreiter M, Course notes of Biological Datasets for Computational Physics, University of Padova (2023).

[20] Manjeet Kumar and others, The Eukaryotic Linear Motif resource: 2022 release, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D497–D508.

[21] Zhang N, Pati D. C-terminus of Sororin interacts with SA2 and regulates sister chromatid cohesion. Cell Cycle. 2015;14(6):820-6.

[22] Zhang N, Pati D. Sororin is a master regulator of sister chromatid cohesion and separation. Cell Cycle. 2012 Jun 1;11(11):2073-83.



(a) *COSMIC.* (b) *BioMuta.* (c) *OncoMX.*

FIG. 6: *Frequency of different types of cancers in the analysed datasets.*

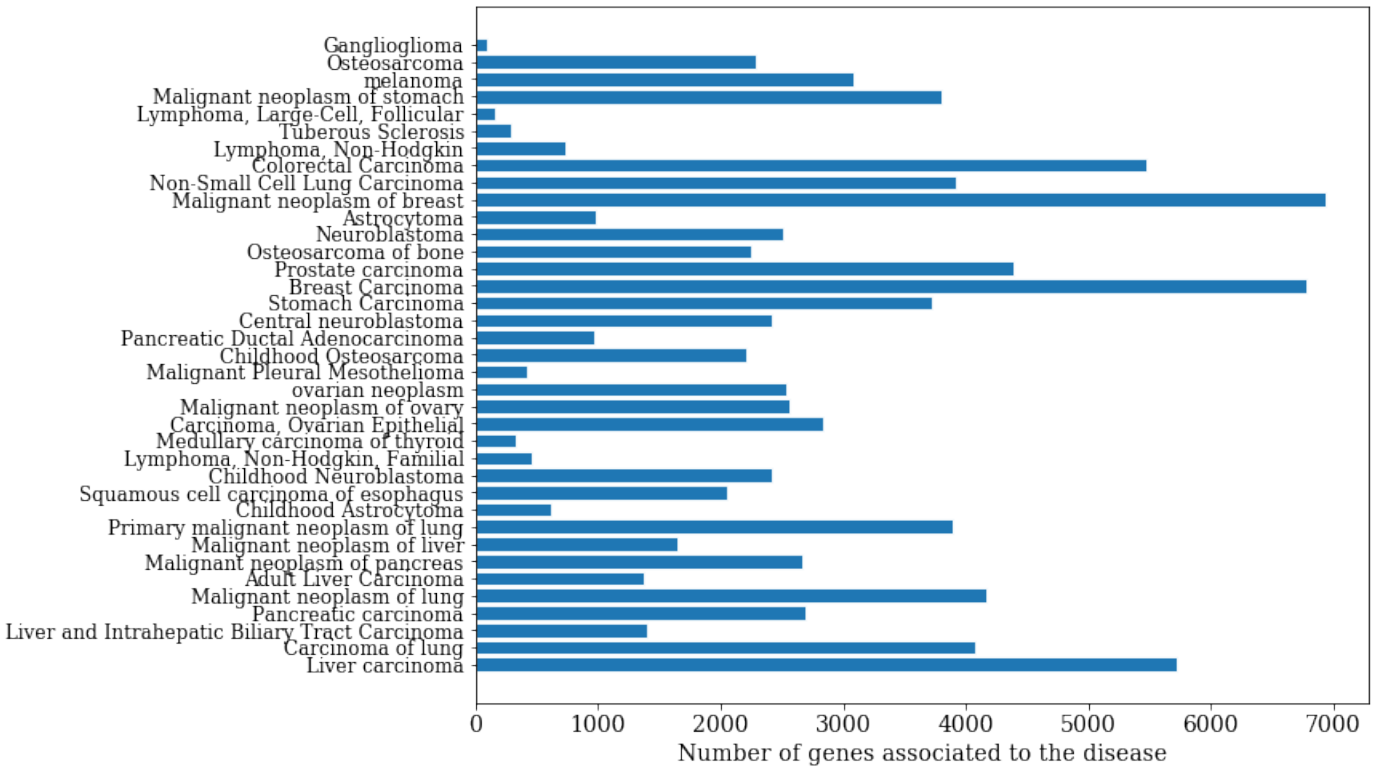


FIG. 7: *Summary of the Gene-Disease Association stored in the BioMuta database, plotting the diseases vs. the number of associated genes.*

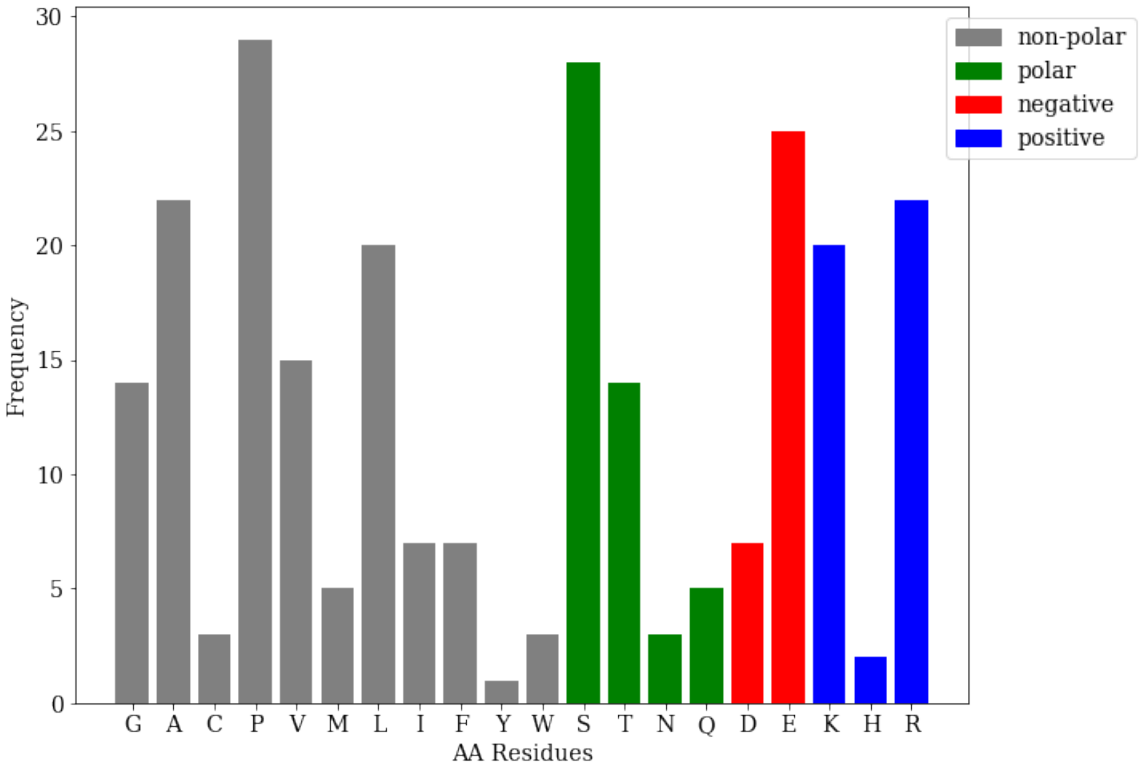


FIG. 8: *Amino acids frequency in Sororin CDCA5 protein, highlighting families/clusters of residues with the same color, resulting in 32 negatively charged residues and 44 positive ones.*

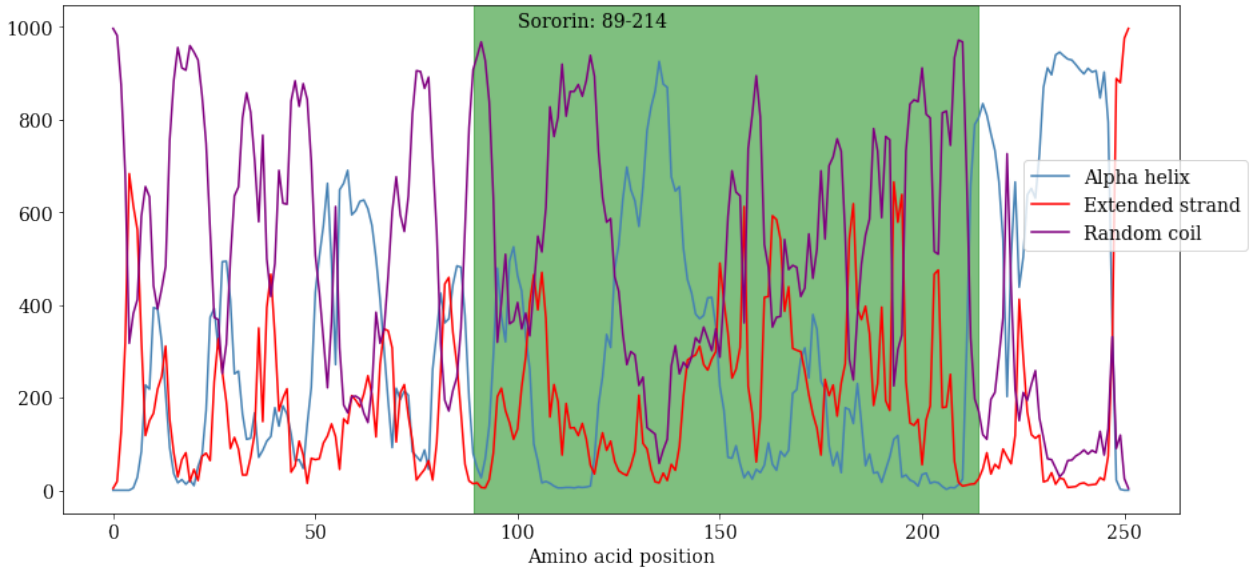


FIG. 9: *Predicted probabilities of the secondary structure of Sororin protein. In green highlighted Sororin subsequence (source: PFAM)*