Human Nervous System Cancer Mutation Analysis from Protein Sequences and Structures

S. P. B. M. Senadheera[1] ,A. R. Weerasinghe[2] and C. R. Wijesinghe[3]

[1,2,3]University of Colombo School of Computing- Sri Lanka

[1]madhubhaniesenadheera@gmail.com, [2]arw@ucsc.cmb.ac.lk, [3]crw@ucsc.cmb.ac.lk

Abstract. According to the central dogma, genomic mutations will have an impact on protein function and structure. We focused on studying the protein sequence and structural changes in human nervous system tumor single nucleotide mutation datasets. We have considered only SNP mutations which alter the protein sequences. Objective of this study is to enrich human neuron system SNP mutation with protein level data. Furthermore, we aimed to provide user friendly visualization for protein level mutations. According to the results from Variant Effect Predictor (VEP), we identified 59148 unique genomic positions in 15574 genes which alter protein sequences. Moreover, 151402 unique protein positions were altered in 49132 proteins. We found clinically significant positions (C9J4N6_132_R>H and P04637_273_R>C) in protein sequence level. Furthermore, in the datasets, Alanine (A) > Threonine (T) (8%) mutation will change the protein position nonpolar (hydrophobic) to polar (hydrophilic) which will significantly change the protein structure. Secondly, we have developed script to connect BISQUE through R for large mutation enrichment. Finally, we designed a protein structure visualization panel for nervous system caner mutations.

Key words: Single nucleotide polymorphism, Protein, Human Nervous system

1. Introduction

All proteins are synthesized based on respective genomic code. This process is named as central dogma [1]. Since genome will determine the protein sequence [1][2], genomic mutations will affect the protein function and structure[3]. Single nucleotide polymorphisms are main type of mutations in genome. Human nervous system cancer data are rarely studied in protein sequence level and protein structure level.

Human nervous system consisted with two parts as central nervous system and peripheral nervous system. We have considered 11 central nervous system caners projects and 4 peripheral nervous system cancers projects for the following study. We focused on studying the protein sequence (amino acid) and structure (topological and 3D structure) changes in nervous system tumor mutation datasets available in cBioPortal[4]. In the cancer research, large dataset handling is required in order to understand the cancer patterns. Furthermore, deep investigation about each entry needs to be considered. We attempted to design a workflow which can handle large dataset while annotating all possible isoforms and predicted structures for each mutations. This will enhance the efficiency and effectiveness genetic mutation investigating process.

2. Objectives

Objective of this study is to map human neuron system SNP mutation into protein level sequence and structural data. Furthermore, we try to provide user friendly visualization for protein level mutations. Finally, we attempt to build a reusable data annotating and visualizing workflow for SNP mutation in protein level. This workflow will be a flexible as a result it can be used for different other cancer types.

3. Methodology

We have considered 15 human nervous system tumor projects for this study. Table 1 shows the datasets we have used for the analysis. Initially we have extracted genomic mutations which changes the protein sequences. For the data filtration, we used Variant Effect Prediction Genome Reference Convertor Tool [5] and Ensemble databases [6]. Secondly, we mapped those genomic level changes with their respective protein level sequences and structures.

We have performed data annotation in two layers. For sequence level data annotation, we have compared four different tools. Ensemble Variant Effect Predictor [7], UNIPROT web tool [8], UNIPROT API and BISQUE tool [9] were used for protein sequence level data annotation comparison. We annotated the protein effect prediction scores for protein alterations via using Polyphen 2[10] and SIFT algorithms[11]. Second data layers was sequence level to protein structure level data annotation. For the above mention step we compared BISQUE [9], PDB genomic mapper [12] and Joseph Marsh Group PDB Mapper tool[13](not available online).

Since we had batch of mutations, we considered the batch processing option in each tool. We have written a R scripts for extract protein level data into each mutation as batches. We have used httr, jsonlite and xml2, data.table libraries for batch processing and data cleaning in R. We have designed a PHP based panel to visualized the PDB structures mapped from the output. We have visualized UNIPROT isoform levels and every predicted protein structure for each mutated protein. This will show whether the mutation is occurred in binding sites, motif or domain. We presented the topological view of the protein sequences therefore it is easy to analyse the impact of the mutations in protein domains and motifs. We integrated the PDBe protein residue interaction component[12] for our visualization panel.

We have identified two major protein alterations in the datasets (IDH1 and TP53 gene mutations) We have compared their wild type structure and mutated protein structures. We investigated the protein structural differences by using RCSB PDB protein structure comparison tool[14].

Designed workflow will enable more user-friendly environment for the cancer researchers in order to analyse protein sequence level and structure level mutation impact on human nervous system single nucleotide polymorphisms. Furthermore followiing approach is reusable for other types of cancers.

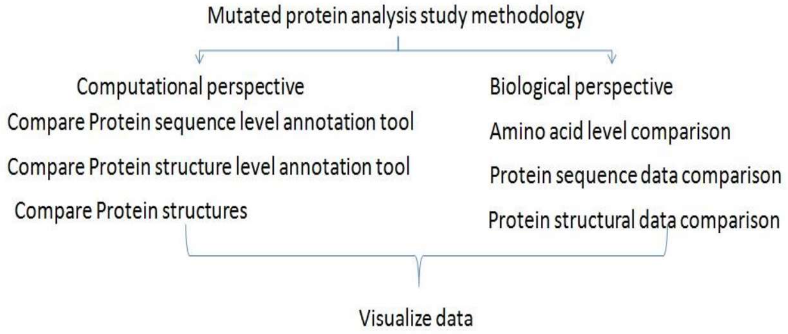

Mutated protein analysis study methodology
Computational perspective
Compare Protein sequence level annotation tool
Compare Protein structure level annotation tool
Compare Protein structures
Biological perspective
Amino acid level comparison
Protein sequence data comparison
Protein structural data comparison
Visualize data

Fig.3. Research methodology for protein level data enrichment in single nucleotide polymorphisms

We designed the research methodology to cover two perspectives in bioinformatics data analysis. This work will contributed to computational and biological domain as shown in figure 1. In computational perspective we analysed 7 protein level data annotation techniques. In biological perspective we have analysed 15 cancer projects which have 1106362 genomic entries and annotated protein sequence level and structural level data. Finally, designed visualization panel combines all effectual tools and annotated data for data mining. Furthermore, we have analysed the amino acid level changes in full dataset. We have analysed protein chemical property changes in the mutated proteins.

Table 2: Dataset details

Type	Project name	Donors	SNP
Peripheral nervous system	mpnst_mskcc	15	3767
Peripheral nervous system	nbl_amc_2012	73	506
Peripheral nervous system	nbl_ucologne_2015	56	818
Peripheral nervous system	nbl_target_2018_pub	372	1035
Central nervous system	pcpg_tcga	184	3823
Central nervous system	past_dkfz_heidelberg_2013	78	217
Central nervous system	lgg_tcga_pan_can_atlas_2018	510	37481
Central nervous system	mbl_sickkids_2016	44	4581
Central nervous system	mbl_icgc	114	1018
Central nervous system	mbl_pcgp	37	536
Central nervous system	mbl_broad_2012	92	1696
Central nervous system	odg_msk_2017	22	229
Central nervous system	gbm_tcga	290	20949
Central nervous system	lgg_tcga	286	9228
Central nervous system	lgg_ucsf_2014	61	14804

4. Results and Discussion

4.1. Sequence level Data Annotation

According to the results, we identified unique 59148 genomic positions in 15574 genes. Moreover, 151402 unique protein positions were altered in 49132 proteins. Initially we have used Variant Effect Predictor tool to extract the protein ids. However, the tool is using ENSP nomenclature as the main ID for proteins which is not compatible with many protein annotating tools. However, it can provide protein mutated position (Amino acid sequence position).

Secondly we used the UNIPROT web tool which has the universal nomenclature for protein sequences and it is compatible with all protein structure databases. However, it could not handle batch processing and it is also not providing the mutated protein position. UNIPROT API can handle batch processing and provide all isoforms of the protein. On the other hand, it only provides the protein sequence start and end positions, not the particular mutation position. BISQUE tool is able to find both UNIPROT ids and its respective protein positions. Furthermore, it provided all the isoforms of the particular mutated protein. This will be an advantage for further analysis on publications regarding the mutations. However, this tool also depends on UNIPROT database. We have considered the Poyphen score, SIFT score and literature publication regarding the mutations. These details were added via using the Variant Effect Prediction tool and UNIPROT database. Furthermore, Protein-protein interaction details and literature regarding mutated proteins were added by using UNIPROT API.

To extract details regarding the protein mutation, we used REST API with R libraries. Data purification had performed after annotating the protein details by using the data.table library in R package. Basic statistics were analysed via using R. We have annotated 1,106,362 entries via using the R script and achieved 99.84% success rate. Figure 2 shows the UNIPROT protein id annotation effectiveness in VEP tool compared to UNIPROT based tools. Even though, VEP tool provides various information regarding a particular mutation, it may not fit for protein isoform identification. Due to the above reason we have merged UNIPROT database for the designed workflow via BISQUE tool and UNIPROT REST API.

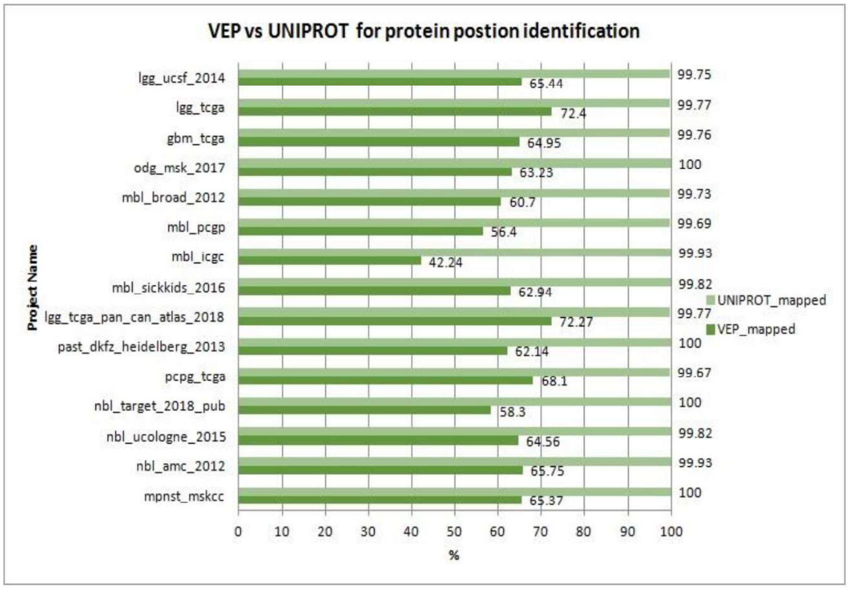


Fig. 4 VEP tool and UNIPROT bases protein position identification protein level data annotation effectiveness

4.2. Basic Statistics

Figure 3 shows the 20 highly mutated genes in the full dataset. IDH1 is the highest mutated gene and TP53 (Tumor Protein 53) is the second highest. IDH1 is vastly related brain cancer types such as glioma and literature supports IDH1 mutations has an impact on tumor progression[15]. Moreover, there are many oncogenes mutated frequently in the dataset such as MUC16 and EGFR. Since mutations are diverse in the dataset (Tumor heterogeneity), in most genes mutated frequency is low.

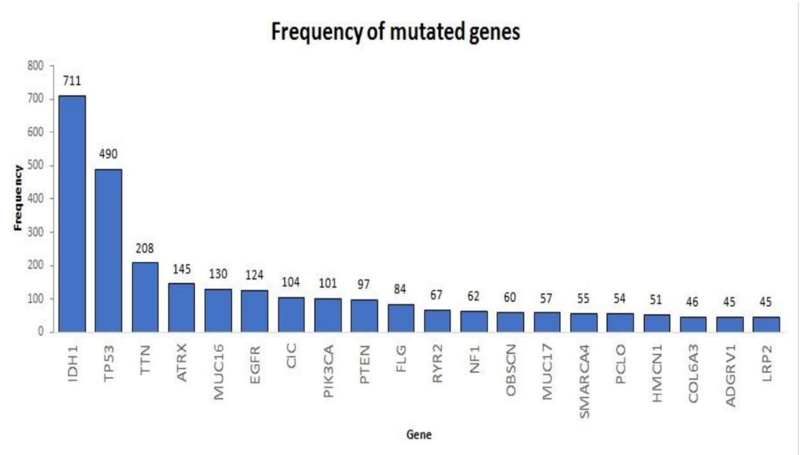


Fig. 5 Mutated genes in the full dataset

Figure 4 shows the isoforms of mutated proteins. Highest mutated isoform proteins are encoded by IDH1 gene and second highest isoform proteins are encoded by TP53 gene. In figure 5, protein position frequencies are shown. Highest mutated protein position (C9J4N6_132_R>H) is clinically significant position according to the drug targeting literature [16]. Patient who carries this particular mutation is more responding to the cancer drugs than the wild type. However, this position is not highly considered in cancer literature in protein impact perspective. There are only two publication available for predicting the impact of the protein mutation and those two have conflicting opinion regarding the mutation.

Second highest mutated protein positions such as P04637_273_R>C is from TP53 which is a common cancer gene. Polyphen and SIFT algorithm predict these positions as pathogenic. Moreover, these position are highly considered in cancer literature.

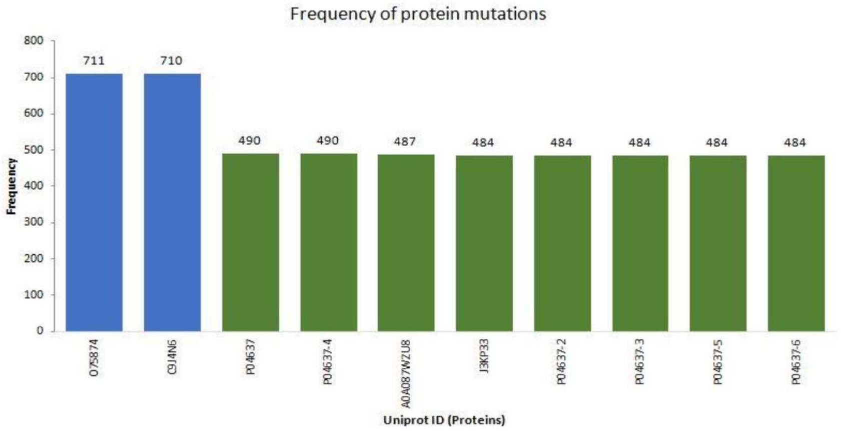


Fig 6 Frequency of mutated proteins in 15 datasets

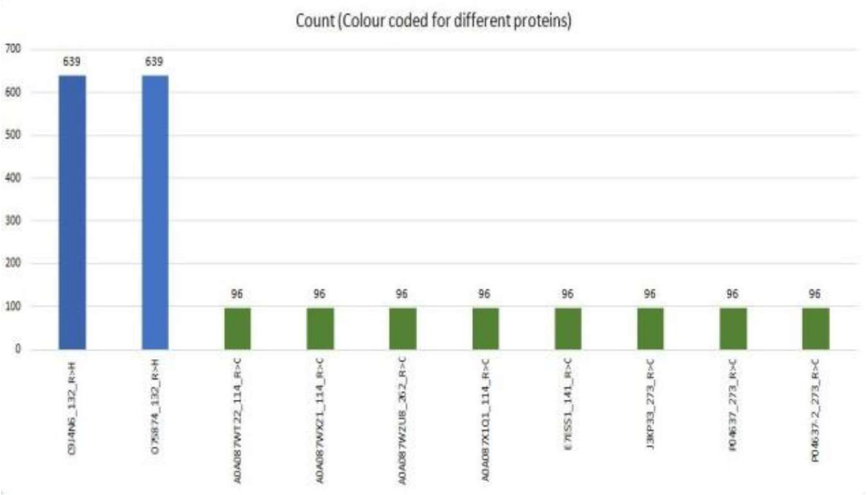


Fig. 7 Frequency of mutated protein positions in full dataset

4.3. Analysis of Amino Acid Changes

Amino acid attributes changes will alter the protein 3D structures and topology. We analysed the amino acid changes in the full dataset. Figure 6 shows the amino acid changes distribution. From the mutated protein, 69.4% are changing the protein attribute due to the mutation.

According to the results, Arginine R> Histidine H change (12%) is the highest alteration possibility. This is a cancer signature which is commonly identified in many cancer types[17]. Both residues are electrically charged (positive and hydrophilic).

However, second highest alteration is changing it's attribute due to the mutation. Alanine (A) > Threonine (T) (8%) mutation will change the protein position nonpolar (hydrophobic) to polar (hydrophilic). These mutations have high chance to change the structure and features of the proteins. Furthermore, mutation Arginine (R) > Cysteine (C) (8%) alteration also changes its' attributes from positively charged to no charge and polar to non polar. However, Alanine (A) > Valine (V) (8%) alteration changes does not change the residue attributes.

Among the original residues, Arginine clearly has the highest relative mutability (36%). According to the literature there is a cancer related biasness in Arginine mutations [17][18]. In figure 7 we drew the wild type residue percentage distribution. Arginine is a positivily charged residue with guanidinium group and forms multiple hydrogen bonds to backbone carbonyl oxygens. Due to hydrogen bonds it helps to create the the structure of the protein [19].

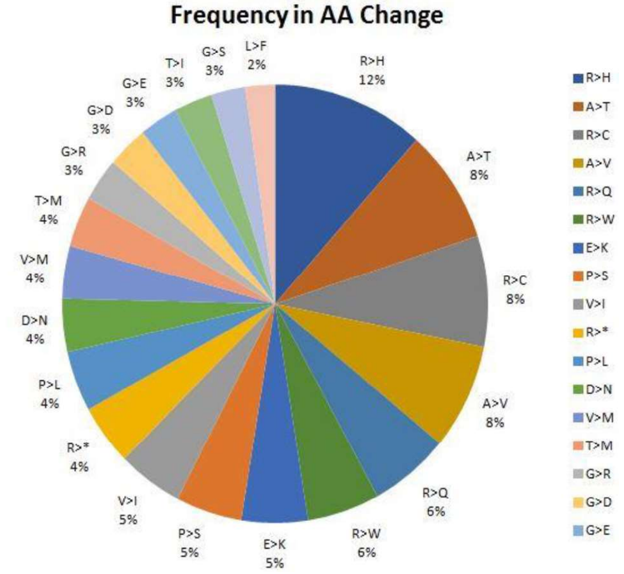


Fig. 8 Amino acid changes in the full dataset

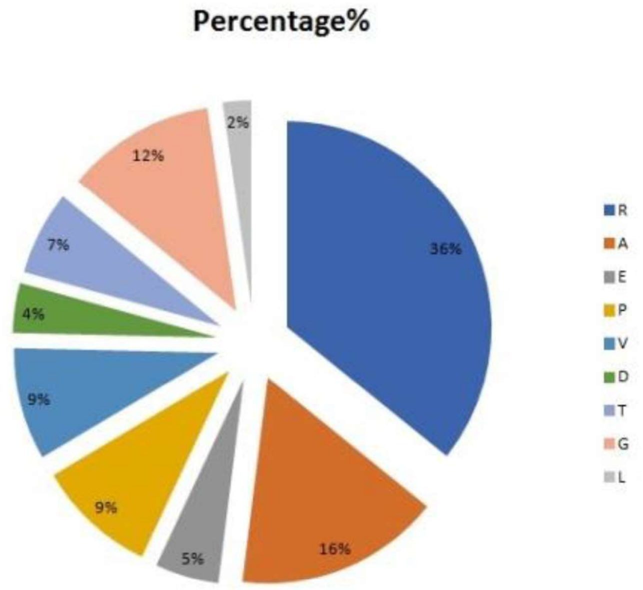


Fig. 9. Wild type residue distribution for the mutated protein

4.4. Structural Level Data Annotation

In the second level data annotation, we used three different tools to identify protein structures of the mutated proteins. PDB genomic mapper tool is a web tool and it provides the PDB chain and position of the mutation. However, it could not handle batch processing. Due to the above reason, it would not be suitable for big data annotation. We validated the other tool's outcome randomly with this tool since it is the direct tool linked to the PDBe database. Since PDBe is the main database which consisted all curated protein structures, other tools refer it for data annotation. BISQUE tool is able to provide the UNIPROT id and PDB structure however, it is not providing the PDB position of the mutations. Since PDB visualization component library is accepting the UNIPROT id and position we used this tool for the protein structure mapping. Joseph Marsh Group PDB mapping tool is able to map PDB structure, position and its attributes by the UNIPROT id and position. However, this tool is not available online. Furthermore, tool is using old version of the database for database annotation. Due to the above restrictions we considered the UNIPROT id, PDB id and UNIPROT position for data visualization. We have annotated 1345 unique proteins with their respective structures. Some UNIPROT sequences do not have PDBe structures. Due to the above reason this R script needs to rerun when PDBe introduce a new data release.

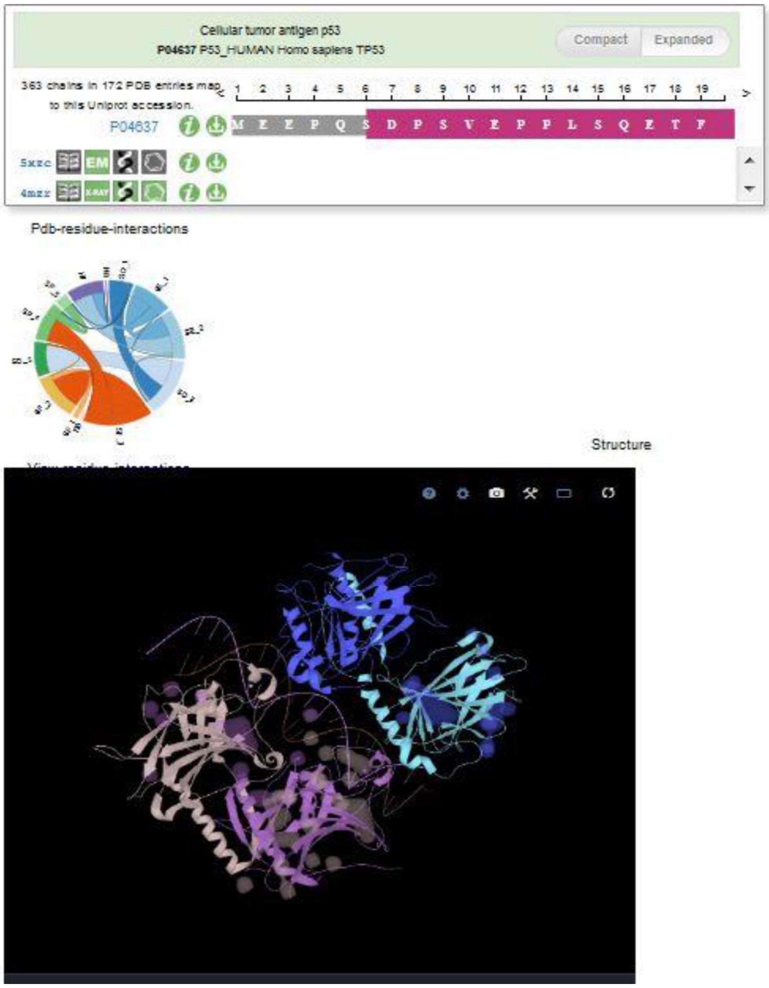
4.5. Protein Data Visualization

In protein sequence and structure visualization, we have used the PDB Topology Viewer, Sequence Feature View, PDB Residue Interactions and LiteMol. Figure 8 and 9 shows the interface of the protein structure visualizing page. LiteMol plugin will enable to download the protein structure and add water molecular interaction visualization. PDB structures are rotatable. This features were added to the workflow in order to make the process easy for oncologists and cancer researchers.

Protein topological graph and 3D visualization is explained in figure 9. By adding this plugin, we tried to improve the user friendliness of the output. Since protein structures are less compared to sequence data this function is only available for well studied proteins. UNIPROT sequence visualization enable user to select the protein structure they prefer based on protein structure quality and protein mutation. If there are protein structures available for wild type and mutated protein, researcher can visualize both proteins and compare the differences occurred due to mutation. By using this panel protein residue location can be identified. Mutations appeared in motifs and domains will be more pathogenic compared to rest since they are in the binding sites.

We have studied highly mutated positions (C9J4N6 132 R>H and P04637 273 R>C) by using visualization panel. We identified wild type protein structure and mutated protein structure for above mentioned mutation. For further investigation we used RCSB PDB structure comparison to validate the mutation results.

In figure 10, we have compared highest mutated protein and it's wild type structure of IDH1 gene (protein isoforms). Mutation occurred in beta sheet of the protein. Yellow colour highlighted areas in figure 10 are the wild type and mutated type protein residue. According to Pfam[20], above mentioned mutation occurred in the DNA binding region. According to the structure comparison in RCSB PDB tool, R>H mutation is structurally equivalent but not same residue.



Cellular tumor antigen p53

P04637 P53_HUMAN Homo sapiens TP53

363 chains in 172 PDB entries map to this Uniprot accession.

Fig. 10 Protein structure visualizing interface

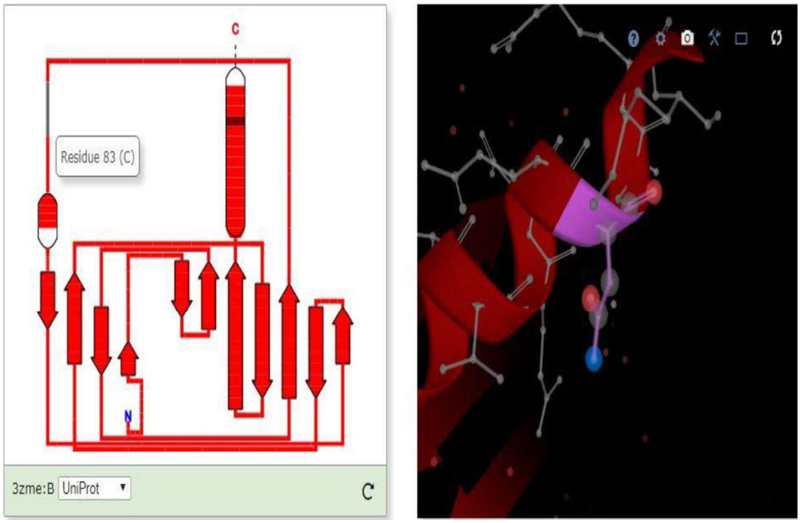
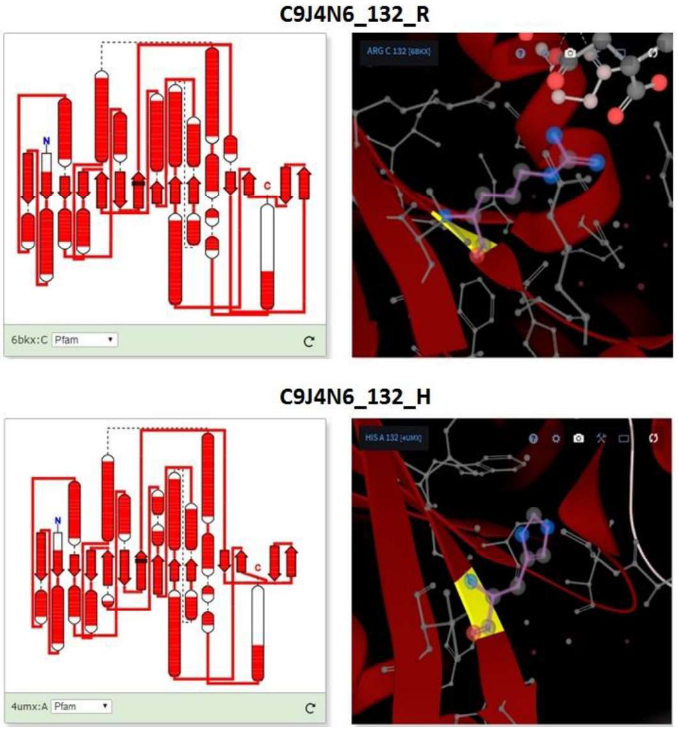


Fig. 11 Visualizing the effect of the residue



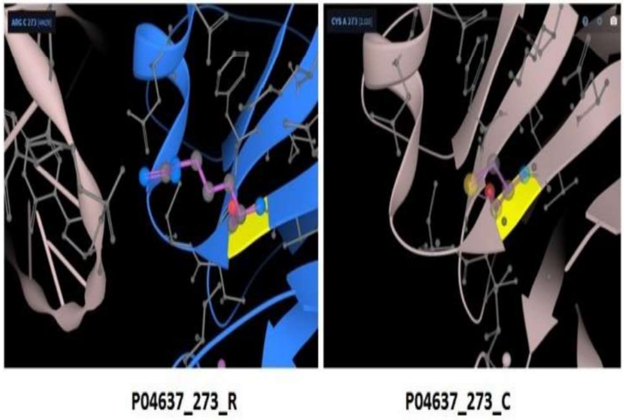
C9J4N6_132_R

C9J4N6_132_H

Fig. 12. C9J4N6_132_R>H mutation visualization in topological and 3d view

In figure 11, TP53 protein mutation is shown. This mutation appeared in beta sheet and also in DNA binding site. However, according to RCSB PDB structure comparison this mutation is structurally equivalent even though, the residue and its' attributes are changed. Since the residue is not similar, the attributes of the protein residue will be altered. Due to the above reason, we can predict that protein binding site function can be altered.

Since this research is mainly based on computational predictions and literature based evidence, these mutated position need to be further investigated in molecular lab in order to verify the results.



P04637_273_R

P04637_273_C

Fig. 13. P04637_273_R>C mutation 3d visualization

5. Conclusions

In biological perspective, we recognized clinically significant positions (C9J4N6 132 R>H and P04637 273 R>C) in protein sequence level. Furthermore, in the dataset, Alanine (A) > Threonine (T) (8%) mutation will change the protein position nonpolar (hydrophobic) to polar (hydrophilic) which will significantly change the protein structure. Arginine (R) > Cysteine (C) (8%) alteration changes its' attributes in positively charged to no charge and polar to non polar. In the study Arginine (R) amino acid is highly mutated among donors.

Secondly, we have developed script to connect BISQUE through R for large mutation list annotation. We compared 7 different approaches for protein level data annotation

developed script achieved 99% success (data loss is less than 1%) in data annotation.
Finally, we designed a protein structure visualization panel for nervous system caner mutations for topological analysis and structural analysis. As future work, we are trying to implement the process as a single workflow and integrate protein structure comparison to the workflow.

6. Acknowledgement

We would like to acknowledge the guidance provided by the European Bioinformatics Institute (United Kingdom) for this study. We would like to thank specially the molecular interaction team for providing the information regarding data and tools for the analysis.

References

1. B. Alberts, A. Hnson, J. Ewis, M. Raff, K. Roberts, and P. Alter, *The Cell*, vol. 40, no. 6. Taylor & Francis Group, LLC, an informa business, 270 Madison Avenue, NewYork NY f 0016, USA, and 2 park Square, Milton park, Abingdon, OXl4 4RN, UK., 2001.
2. K. a. Hoadley *et al.*, “Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin,” *Cell*, vol. 158, no. 4, pp. 929–944, 2014.
3. I. Rehman and S. Botelho, *Biochemistry, Secondary Protein Structure*. StatPearls Publishing, 2019.
4. J. Gao *et al.*, “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal,” *Sci. Signal.*, vol. 6, no. 269, pp. pl1–pl1, Apr. 2013.
5. W. McLaren *et al.*, “The Ensembl Variant Effect Predictor,” *Genome Biol.*, vol. 17, no. 1, p. 122, Dec. 2016.
6. “Ensembl genome browser 98.” [Online]. Available: https://asia.ensembl.org/index.html. [Accessed: 16-Nov-2019].
7. W. McLaren *et al.*, “The Ensembl Variant Effect Predictor,” *Genome Biol.*, vol. 17, no. 1, p. 122, Dec. 2016.
8. “UniProt.” [Online]. Available: https://www.uniprot.org/. [Accessed: 17-May-2019].
9. M. J. Meyer, P. Geske, and H. Yu, “BISQUE: locus- and variant-specific conversion of genomic, transcriptomic and proteomic database identifiers,” *Bioinformatics*, vol. 32, no. 10, pp. 1598–1600, May 2016.
10. I. A. Adzhubei *et al.*, “A method and server for predicting damaging missense mutations,” *Nat. Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010.
11. N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, “SIFT web server: predicting effects of amino acid substitutions on proteins,” *Nucleic Acids Res.*, vol. 40, no. W1, pp. W452–W457, Jul. 2012.
12. “PDB Component Library &gt; Documentation.” [Online]. Available: https://www.ebi.ac.uk/pdbe/pdb-component-library/doc.html#a_topologyViewer. [Accessed: 21-May-2019].
13. “Joe Marsh Research Group | The University of Edinburgh.” [Online]. Available: https://www.ed.ac.uk/mrc-human-genetics-unit/research/marsh-group. [Accessed: 16-Nov-2019].
14. H. M. Berman, “The Protein Data Bank / Biopython,” *Presentation*, vol. 28, no. 1, pp. 235–242, 2000.
15. L. Deng *et al.*, “Association between IDH1/2 mutations and brain glioma grade.,” *Oncol. Lett.*, vol. 16, no. 4, pp. 5405–5409, Oct. 2018.
16. H. Yan *et al.*, “IDH1 and IDH2 mutations in gliomas.,” *N. Engl. J. Med.*, vol. 360, no. 8, pp. 765–73, Feb. 2009.
17. V. Tsuber, Y. Kadamov, L. Brautigam, U. W. Berglund, and T. Helleday, “Mutations in Cancer Cause Gain of Cysteine, Histidine, and Tryptophan at the Expense of a Net Loss of Arginine on the Proteome Level.,” *Biomolecules*, vol. 7, no. 3, 2017.
18. S. Khan and M. Vihinen, “Spectrum of disease-causing mutations in protein secondary structures.,” *BMC Struct. Biol.*, vol. 7, p. 56, Aug. 2007.
19. C. L. Borders *et al.*, “A structural role for arginine in proteins: multiple hydrogen bonds to backbone carbonyl oxygens.,” *Protein Sci.*, vol. 3, no. 4, pp. 541–8, Apr. 1994.
20. S. El-Gebali *et al.*, “The Pfam protein families database in 2019,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D427–D432, Jan. 2019.