Modeling TP53 Substitution Mutation Frequencies in Various Human Tissues Using a Markov Model

Carly Kingsbury & Mallory D’Antonio
Cape Cod Community College

INTRODUCTION

The TP53 gene contains the genetic code of instructions for the p53 protein. This protein is known as a tumor suppressor. When functioning properly, if it detects damaged cells it will stop their division cycle while initiating the death of those damaged cells, otherwise known as apoptosis. If p53 determines that the cell can be repaired it will signal other genes to activate and repair the damaged cells. Inactivation of this gene allows damaged cells to continually reproduce increasing the chance of a tumor to form.

A mutation in the TP53 gene can be inherited genetically or is acquired. If it is passed on by a parent, the mutation is known as the rare genetic disorder, Li-Fraumeni syndrome. Inheriting this mutation genetically is less common than a somatic mutation, but is still possible with a family history of the mutation.

Base substitution mutations are one of the most common mutation types to occur in the human TP53 gene, with missense substitutions accounting for approximately 61% of all observed TP53 mutations. The rate and probability of the occurrence of each base substitution can be modeled with a Markov or stochastic matrix, with each entry in the matrix representing the probability of a specific base substitution from an ancestral DNA strand to a descendent strand occurring after a single time step. In this manner, a universal model for base substitution probability in the TP53 gene can be constructed from a reference sequence and known substitution rates. The focus of this investigation is on examining the substitution rates in TP53 sequence samples from various human tissues in order to determine if the derived Markov matrix accurately models substitution mutation rates in TP53 across different tissue types.

<u>Research Question</u>: To what extent can the frequencies of base substitution mutations in TP53 gene sequence samples from various human tissues be accurately modeled by a Markov model?

METHODOLOGY

In order to derive the Markov matrix used in this investigation, the reference sequence NM_000546 for the TP53 gene was first downloaded from the National Center for Biotechnology Information and viewed in MATLAB using the Sequence Viewer app in order to determine the exact number of adenosine, cytosine, guanosine and thymine bases present in TP53. This information was then used in conjunction with data from COSMIC outlining the percentages of TP53 samples in which each type of base substitution was observed in order to create a Markov matrix of base substitution probabilities in the TP53 gene. The substitution rate predictions of this Markov matrix were then compared to the actual percentage of samples in which each mutation type occurred in various cancerous human tissues in order to investigate the consistency of the mutation pattern across different tissue types.

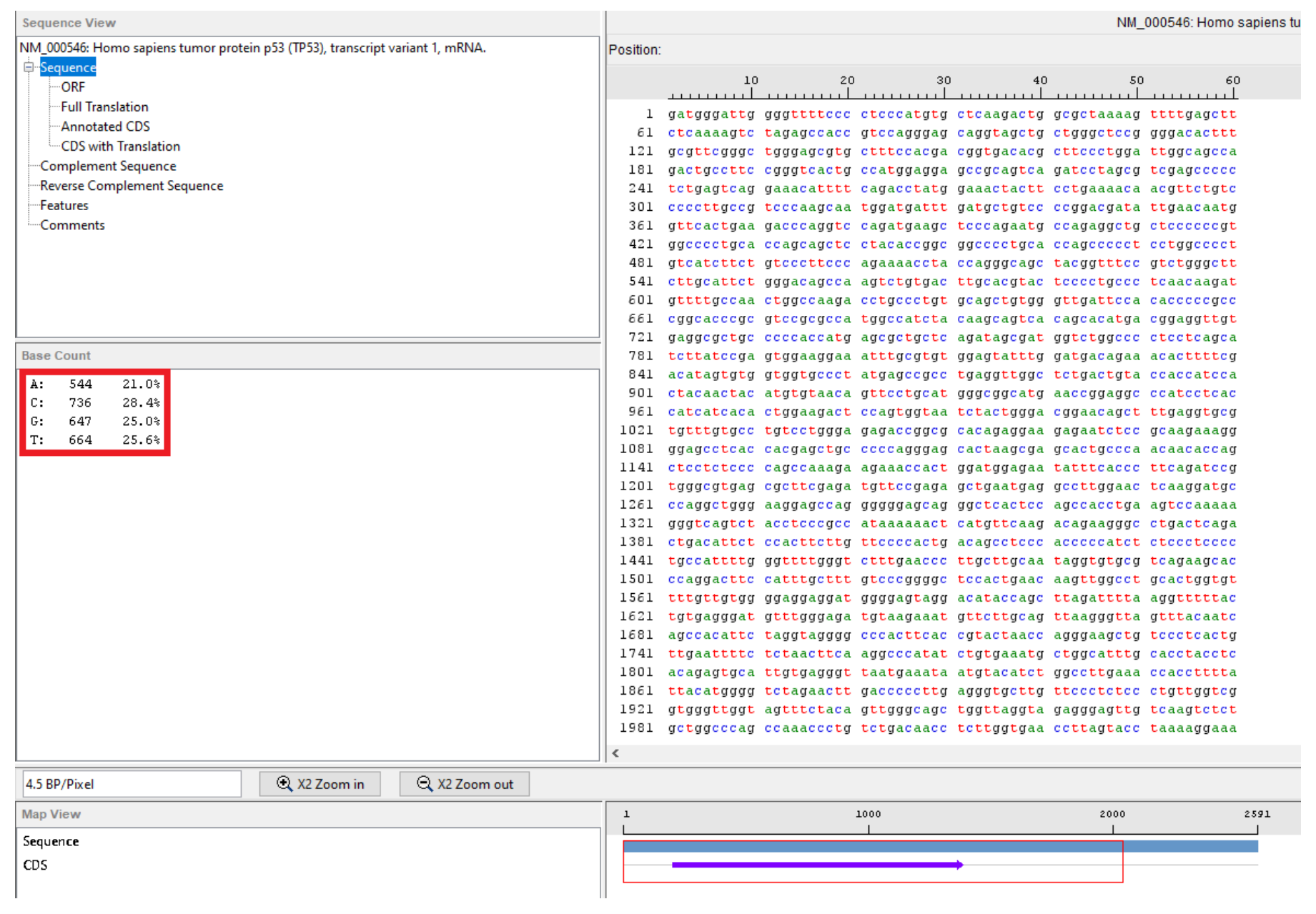
Figure 1: Formula for a Markov Model of Base Substitution Mutation

$$
M\mathbf{p}_0 = \begin{pmatrix} \mathcal{P}_{A|A} & \mathcal{P}_{A|G} & \mathcal{P}_{A|C} & \mathcal{P}_{A|T} \\ \mathcal{P}_{G|A} & \mathcal{P}_{G|G} & \mathcal{P}_{G|C} & \mathcal{P}_{G|T} \\ \mathcal{P}_{C|A} & \mathcal{P}_{C|G} & \mathcal{P}_{C|C} & \mathcal{P}_{C|T} \\ \mathcal{P}_{T|A} & \mathcal{P}_{T|G} & \mathcal{P}_{T|C} & \mathcal{P}_{T|T} \end{pmatrix} \begin{pmatrix} \mathcal{P}_A \\ \mathcal{P}_G \\ \mathcal{P}_C \\ \mathcal{P}_T \end{pmatrix}
= \begin{pmatrix} \mathcal{P}_{A|A}\mathcal{P}_A + \mathcal{P}_{A|G}\mathcal{P}_G + \mathcal{P}_{A|C}\mathcal{P}_C + \mathcal{P}_{A|T}\mathcal{P}_T \\ \mathcal{P}_{C|A}\mathcal{P}_A + \mathcal{P}_{C|G}\mathcal{P}_G + \mathcal{P}_{C|C}\mathcal{P}_C + \mathcal{P}_{C|T}\mathcal{P}_T \\ \mathcal{P}_{G|A}\mathcal{P}_A + \mathcal{P}_{G|G}\mathcal{P}_G + \mathcal{P}_{G|C}\mathcal{P}_C + \mathcal{P}_{G|T}\mathcal{P}_T \\ \mathcal{P}_{T|A}\mathcal{P}_A + \mathcal{P}_{T|G}\mathcal{P}_G + \mathcal{P}_{T|C}\mathcal{P}_C + \mathcal{P}_{T|T}\mathcal{P}_T \end{pmatrix}
$$

Figure 2: Derived Markov Matrix for Base Substitution Mutation Probabilities in TP53 Gene

S1/S0	A	C	G	T
A	0.8603	0.0313	0.2921	0.0286
C	0.0147	0.6725	0.0464	0.0377
G	0.0901	0.0394	0.5069	0.0331
T	0.0349	0.2568	0.1546	0.9006

Figure 3: NM_000546 TP53 Reference Sequence as Viewed in MATLAB Sequence Viewer



Sequence View
NM_000546: Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA.
- Sequence
  - ORF
  - Full Translation
  - Annotated CDS
  - CDS with Translation
- Complement Sequence
- Reverse Complement Sequence
- Features
- Comments

Base Count

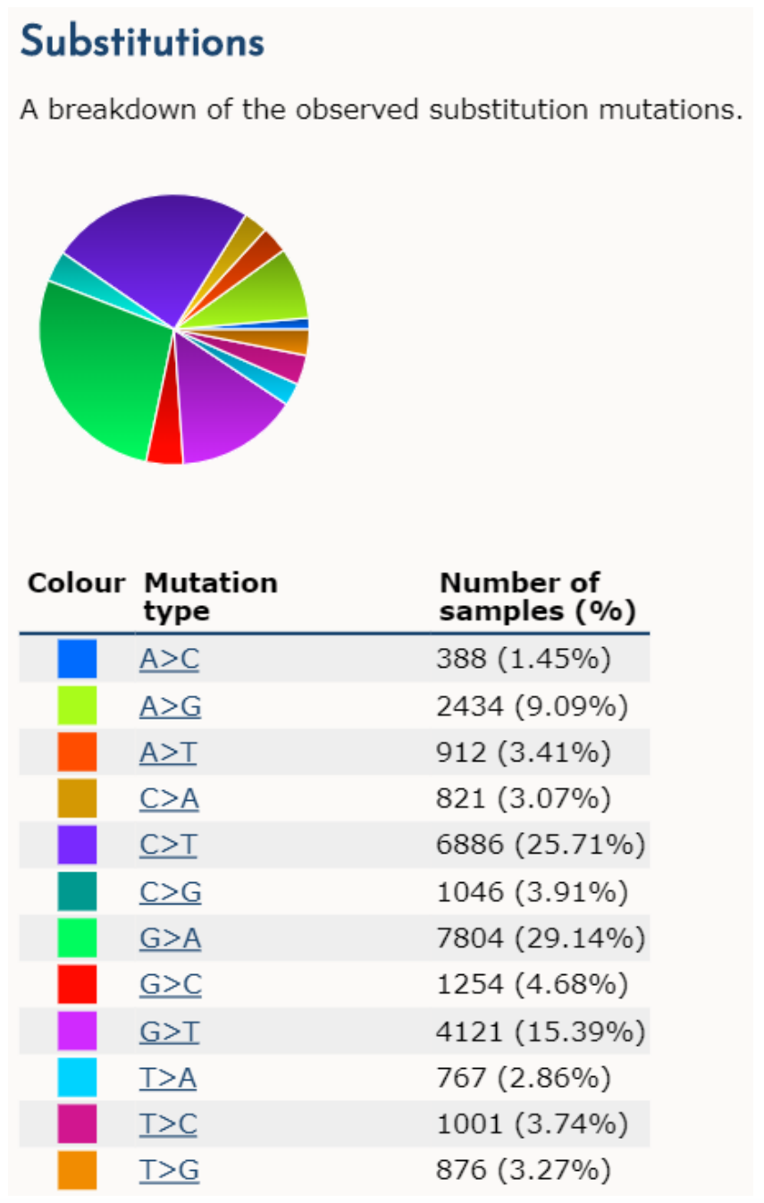
```
A:   544   21.0%
C:   736   28.4%
G:   647   25.0%
T:   664   25.6%
```

NM_000546: Homo sapiens tu

Position:

```
        10         20         30         40         50         60
   1 gatgggattg gggttttccc ctcccatgtg ctcaagactg gcgctaaaag ttttgagctt
  61 ctcaaaagtc tagagccacc gtccagggag caggtagctg ctgggctccg gggacacttt
 121 gcgttcgggc tgggagcgtg ctttccacga cggtgacacg cttccctgga ttggcagcca
 181 gactgccttc cgggtcactg ccatggagga gccgcagtca gatcctagcg tcgagccccc
 241 tctgagtcag gaaacatttt cagacctatg gaaactactt cctgaaaaca acgttctgtc
 301 ccccttgccg tcccaagcaa tggatgattt gatgctgtcc ccggacgata ttgaacaatg
 361 gttcactgaa gacccaggtc cagatgaagc tcccagaatg ccagaggctg ctccccccgt
 421 ggcccctgca ccagcagctc ctacaccggc ggcccctgca ccagccccct cctggcccct
 481 gtcatcttct gtcccttccc agaaaaccta ccagggcagc tacggtttcc gtctgggctt
 541 cttgcattct gggacagcca agtctgtgac ttgcacgtac tcccctgccc tcaacaagat
 601 gttttgccaa ctggccaaga cctgccctgt gcagctgtgg gttgattcca caccccgccc
 661 cggcacccgc gtccgcgcca tggccatcta caagcagtca cagcacatga cggaggttgt
 721 gaggcgctgc ccccaccatg agcgctgctc agatagcgat ggtctggccc ctcctcagca
 781 tcttatccga gtggaaggaa atttgcgtgt ggagtatttg gatgacagaa acacttttcg
 841 acatagtgtg gtggtgccct atgagccgcc tgaggttggc tctgactgta ccaccatcca
 901 ctacaactac atgtgtaaca gttcctgcat gggcggcatg aaccggaggc ccatcctcac
 961 catcatcaca ctggaagact ccagtggtaa tctactggga cggaacagct ttgaggtgcg
1021 tgtttgtgcc tgtcctggga gagaccggcg cacagaggaa gagaatctcc gcaagaaagg
1081 ggagcctcac cacgagctgc ccccagggag cactaagcga gcactgccca acaacaccag
1141 ctcctctccc cagccaaaga agaaaccact ggatggagaa tatttcaccc ttcagatccg
1201 tgggcgtgag cgcttcgaga tgttccgaga gctgaatgag gccttggaac tcaaggatgc
1261 ccaggctggg aaggagccag gggggagcag ggctcactcc agccacctga agtccaaaaa
1321 gggtcagtct acctcccgcc ataaaaaact catgttcaag acagaagggc ctgactcaga
1381 ctgacattct ccacttcttg ttccccactg acagcctccc acccccatct ctccctcccc
1441 tgccattttg ggttttgggt ctttgaaccc ttgcttgcaa taggtgtgcg tcagaagcac
1501 ccaggacttc catttgcttt gtcccggggc tccactgaac aagttggcct gcactggtgt
1561 tttgttgtgg ggaggaggat ggggagtagg acataccagc ttagatttta aggtttttac
1621 tgtgagggat gtttgggaga tgtaagaaat gttcttgcag ttaagggtta gtttacaatc
1681 agccacattc taggtagggg cccacttcac cgtactaacc agggaagctg tccctcactg
1741 ttgaattttc tctaacttca aggcccatat ctgtgaaatg ctggcatttg cacctacctc
1801 acagagtgca ttgtgagggt taatgaaata atgtacatct ggccttgaaa ccacctttta
1861 ttacatgggg tctagaactt gacccccttg agggtgcttg ttccctctcc ctgttggtcg
1921 gtgggttggt agtttctaca gttgggcagc tggttaggta gagggagttg tcaagtctct
1981 gctggcccag ccaaaccctg tctgacaacc tcttggtgaa ccttagtacc taaaaggaaa
```

Figure 4: Percentages of COSMIC TP53 Samples in Which Each Substitution Mutation Type Was Observed



Substitutions

A breakdown of the observed substitution mutations.

Colour	Mutation type	Number of samples (%)
🟦	A>C	388 (1.45%)
🟩	A>G	2434 (9.09%)
🟧	A>T	912 (3.41%)
🟨	C>A	821 (3.07%)
🟪	C>T	6886 (25.71%)
🟦	C>G	1046 (3.91%)
🟩	G>A	7804 (29.14%)
🟥	G>C	1254 (4.68%)
🟪	G>T	4121 (15.39%)
🟦	T>A	767 (2.86%)
🟪	T>C	1001 (3.74%)
🟧	T>G	876 (3.27%)

RESULTS

Key:

Lower frequency than expected
Higher frequency than expected

Margin = 50% lower or higher than expected value

Breast
Table 1: Base Substitution Frequencies in TP53 Samples from Breast Tissue

S1/S0	A	C	G	T
A	0.8517	0.0285	0.2867	0.0354
C	0.0157	0.6958	0.0537	0.0533
G	0.1063	0.0409	0.5544	0.0435
T	0.0263	0.2348	0.1052	0.8678

Esophagus
Table 2: Base Substitution Frequencies in TP53 Samples from Esophagus Tissue

S1/S0	A	C	G	T
A	0.8485	0.0381	0.2940	0.0408
C	0.0085	0.6632	0.0281	0.0302
G	0.0943	0.0302	0.5190	0.0339
T	0.0487	0.2685	0.1589	0.8951

Prostate
Table 3: Base Substitution Frequencies in TP53 Samples from Prostate Tissue

S1/S0	A	C	G	T
A	0.8317	0.0277	0.2594	0.0356
C	0.0198	0.7050	0.0238	0.0554
G	0.1208	0.0495	0.5683	0.0475
T	0.0277	0.2178	0.1485	0.8615

Kidney
Table 4: Base Substitution Frequencies in TP53 Samples from Kidney Tissue

S1/S0	A	C	G	T
A	0.9323	0.0391	0.2598	0.0071
C	0.0107	0.6228	0.1174	0.0214
G	0.0463	0.0498	0.4555	0.0427
T	0.0107	0.2883	0.1673	0.9288

Thyroid
Table 5: Base Substitution Frequencies in TP53 Samples from Thyroid Tissue

S1/S0	A	C	G	T
A	0.8534	0.0263	0.3421	0.0188
C	0.0150	0.6428	0.0639	0.0113
G	0.1128	0.0677	0.4662	0.0038
T	0.0188	0.2632	0.1278	0.9661

Adrenal Gland
Table 6: Base Substitution Frequencies in TP53 Samples from Adrenal Gland Tissue

S1/S0	A	C	G	T
A	0.8876	0.0449	0.2472	0.0112
C	0.0000	0.6517	0.0899	0.0112
G	0.0562	0.0225	0.4831	0.0449
T	0.0562	0.2809	0.1798	0.9327

Urinary Tract
Table 7: Base Substitution Frequencies in TP53 Samples from Urinary Tract Tissue

S1/S0	A	C	G	T
A	0.8346	0.0269	0.3661	0.0143
C	0.0067	0.7221	0.1184	0.0193
G	0.0932	0.0386	0.3954	0.0160
T	0.0655	0.2124	0.1201	0.9504

Liver
Table 8: Base Substitution Frequencies in TP53 Samples from Liver Tissue

S1/S0	A	C	G	T
A	0.8403	0.0367	0.1245	0.0391
C	0.0160	0.8380	0.0415	0.0359
G	0.0910	0.0287	0.3959	0.0319
T	0.0527	0.0966	0.4381	0.8931

Lung
Table 9: Base Substitution Frequencies in TP53 Samples from Lung Tissue

S1/S0	A	C	G	T
A	0.8231	0.0257	0.1610	0.0181
C	0.0194	0.7813	0.0788	0.0171
G	0.1003	0.0460	0.4163	0.0260
T	0.0572	0.1470	0.3439	0.9388

Cervix
Table 10: Base Substitution Frequencies in TP53 Samples from Cervix Tissue

S1/S0	A	C	G	T
A	0.8929	0.0476	0.3452	0.0000
C	0.0238	0.6786	0.0952	0.0119
G	0.0595	0.0476	0.4406	0.0238
T	0.0238	0.2262	0.1190	0.9643

CONCLUSION

It can be concluded that the Markov model derived for this investigation was highly accurate in predicting the frequencies of base substitution mutations in the TP53 gene across samples collected from ten different human tissues. As such, the model has demonstrated that the TP53 gene follows a relatively consistent substitution mutation pattern regardless of tissue type. Furthermore, the derived Markov matrix most accurately modeled TP53 substitution mutation in breast, esophagus and prostate tissue and least accurately in lung and cervix tissue. This result was somewhat unexpected due to the fact that TP53 samples from cancerous esophagus cells have a greater than 80% frequency of mutation and as such the matrix for esophagus substitution frequencies was expected to deviate from the model in the form of increased substitution probabilities. Discrepancies between expected and actual data can be attributed primarily to the involvement of cancer-causing mutations in genes other than TP53 which result in tissue-specific cancers as well as the varying sample sizes of TP53 data from different human tissues in the COSMIC database.

Figure 5: TP53 Alteration Frequency in Various Human Tissues

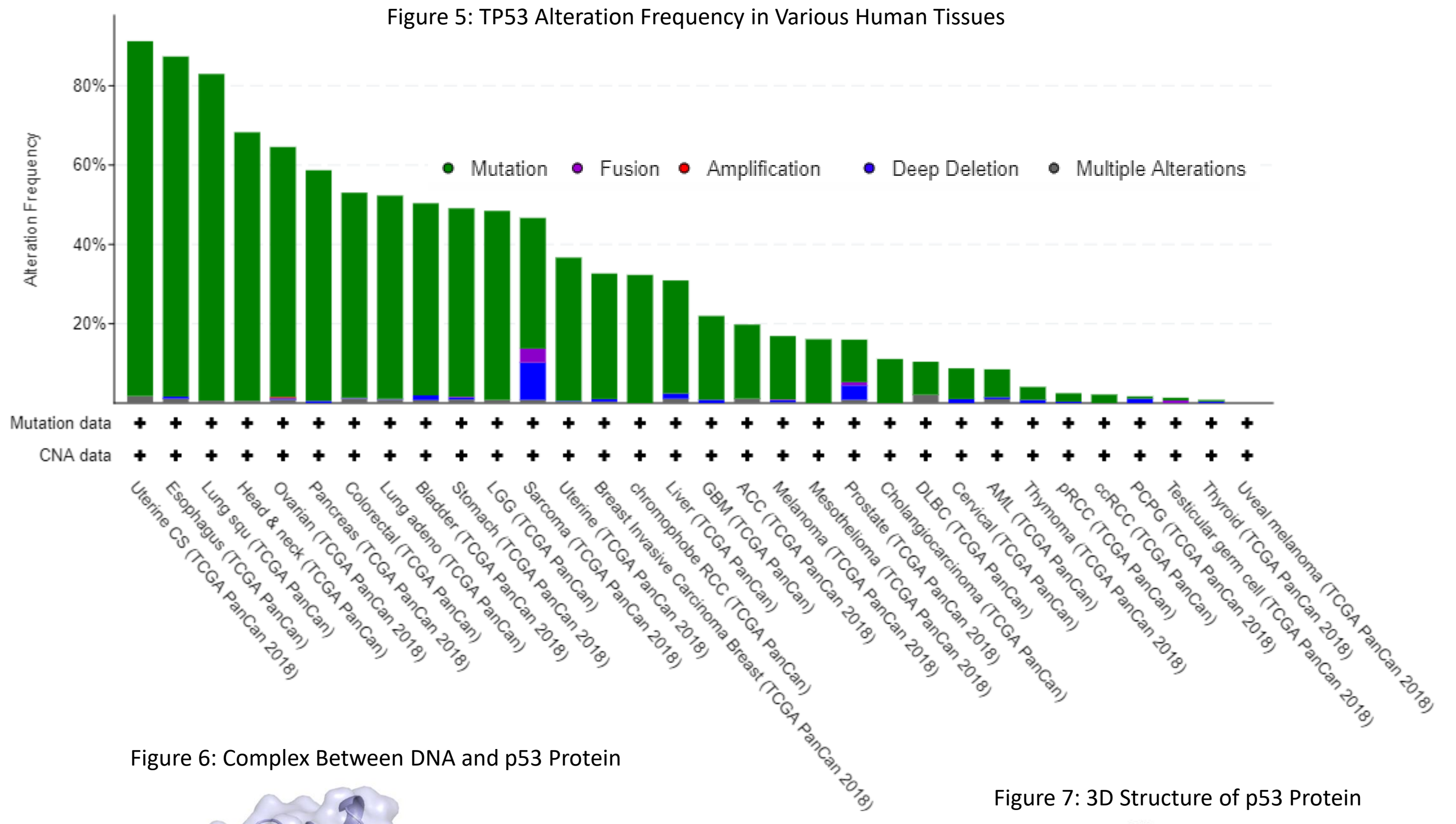


Figure 6: Complex Between DNA and p53 Protein

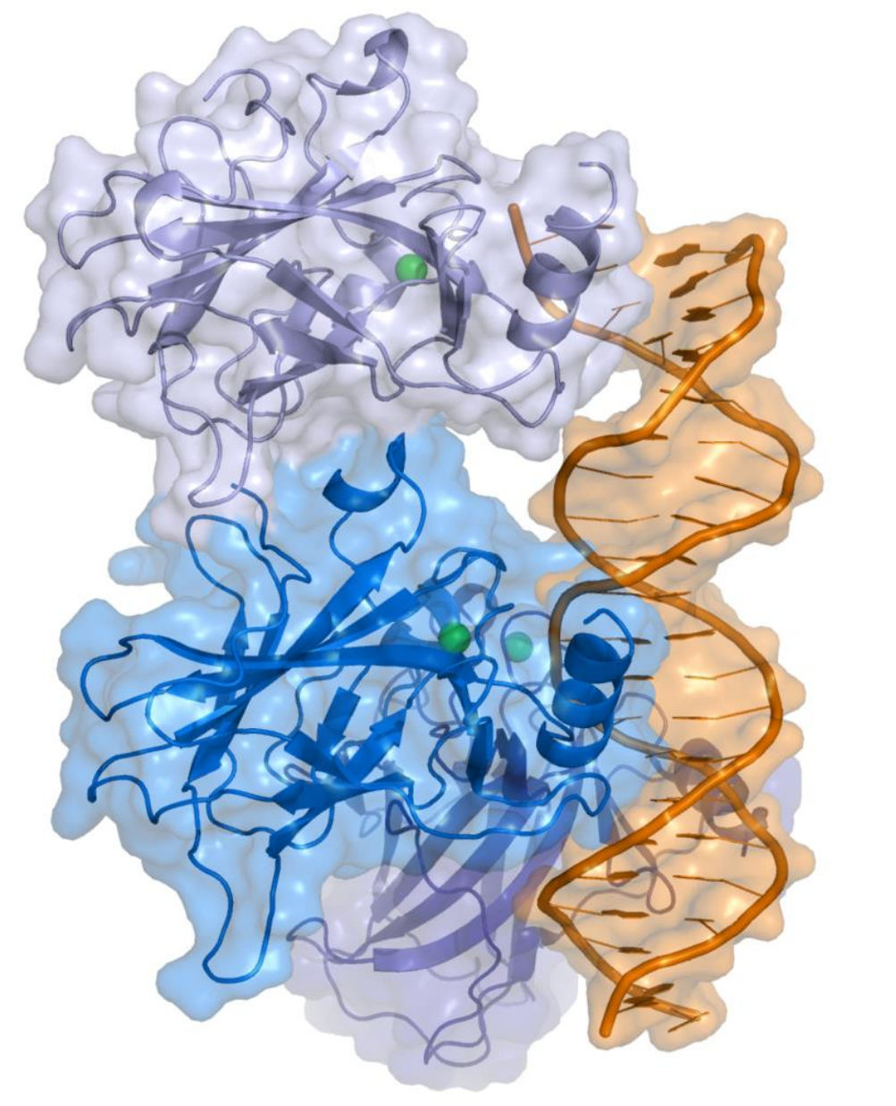


Figure 7: 3D Structure of p53 Protein

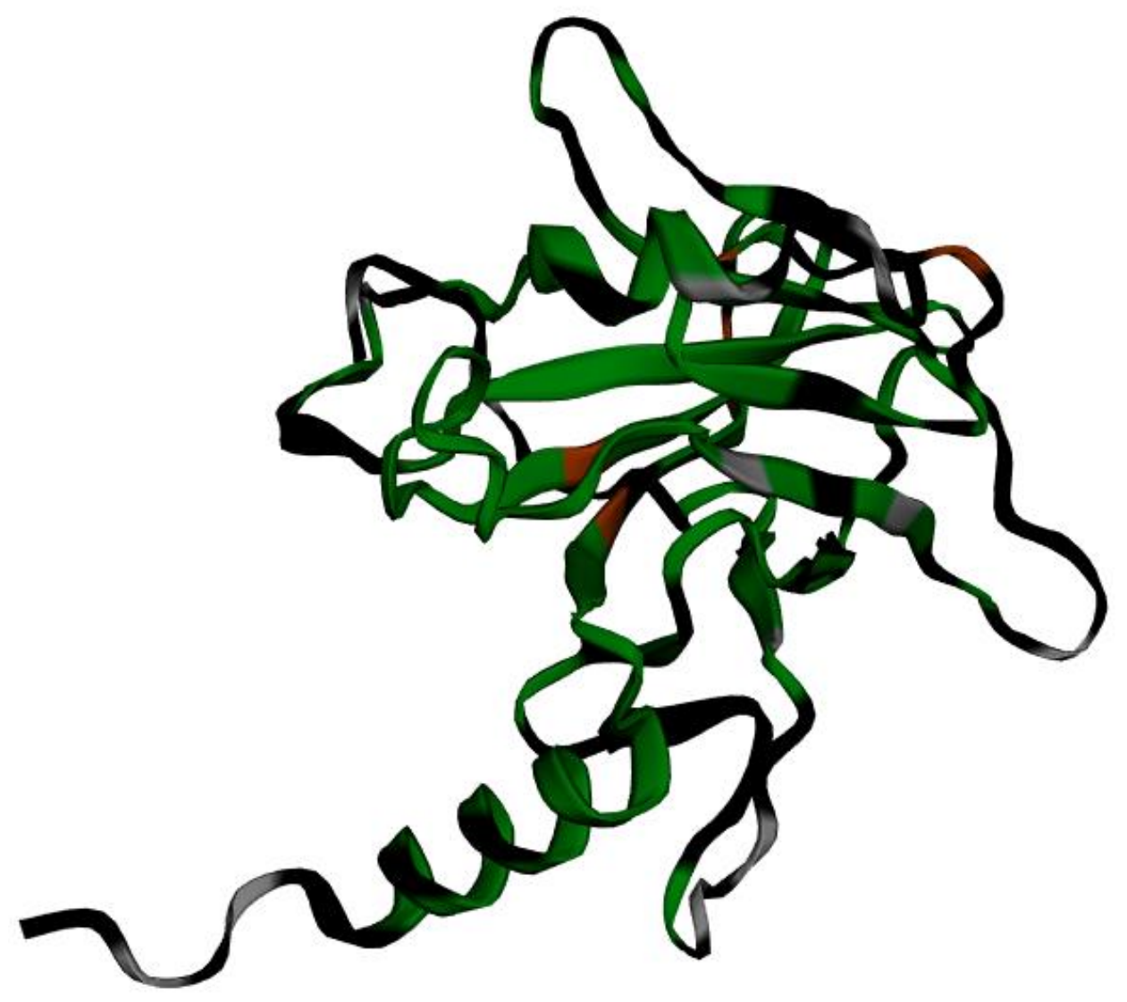
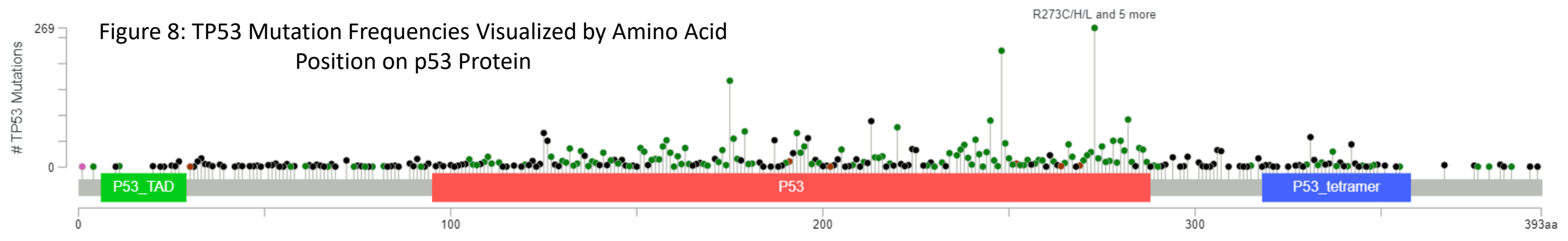


Figure 8: TP53 Mutation Frequencies Visualized by Amino Acid Position on p53 Protein



REFERENCES

Splettstoesser, Thomas. “P53.” *p53*, 5 Oct. 2006, en.wikipedia.org/wiki/File:P53.png.

“4.4 Matrix Models of Base Substitution.” *Mathematical Models in Biology: an Introduction*, by Elizabeth Spencer Allman and John A. Rhodes, Cambridge University Press, 2010.

“TP53 Genetic Test: MedlinePlus Lab Test Information.” *MedlinePlus*, U.S. National Library of Medicine, 4 Feb. 2019, medlineplus.gov/lab-tests/tp53-genetic-test/.

“TP53 Gene - Genetics Home Reference - NIH.” *U.S. National Library of Medicine*, National Institutes of Health, ghr.nlm.nih.gov/gene/TP53.

“TP53.” *CBioPortal for Cancer Genomics*, www.cbioportal.org/results/mutations?cancer_study_list=5c8a7d55e4b046111fee2296&case_set_id=all&gene_list=TP53.

“TP53 Gene - COSMIC.” *TP53 Gene - Somatic Mutations in Cancer*, COSMIC - Catalogue of Somatic Mutations in Cancer, 5 Sept. 2019, cancer.sanger.ac.uk/cosmic/gene/analysis?ln=TP53.

“p53 Mutations in 10,000 Cancer Patients Shed New Light on Gene's Function.” *ScienceDaily*, ScienceDaily, 30 July 2019, www.sciencedaily.com/releases/2019/07/190730141834.htm.

Carly Kingsbury: MATLAB programming and mathematical modeling
Mallory D’Antonio: Research of the biological context of the p53 gene and the cancer-causing effects of base substitution mutations.