Protein Engineering and Computational Analysis of Enzymes to Predict Cancer Mutations

Nolan Sarmiento

ABSTRACT

Cancer is a dangerous disease that can manifest in multiple forms and areas within the body. Due to its multifaceted nature and tenacity, cancer prevention and treatment methods remain challenging. However, developing methods are ongoing that may help with fighting cancer by engaging in preventing growth at early stages. Protein engineering and enzyme design is a promising field that allows proteins to be strategically edited and designed to target cancers more efficiently. Computational methods allow us to obtain a better understanding of how these proteins can be engineered to fight a disease like cancer. Computational software offers data on protein alignments over multiple species, theoretical protein structures, and predicting co-evolving enzyme pairs, as well as making inferences on how mutations affect residues. Here, I aimed to examine computational protein engineering by exploring predictive software that is publicly accessible, to gain insight into their abilities and draw conclusions on the relation between cancer malignancies and the computational results. Overall, computational analysis and understanding enzyme design can have exciting implications for the world of medicine and allow for robust means of treating diseases like cancer.

INTRODUCTION

The field of protein engineering and enzyme design has experienced a transformative surge driven by remarkable advancements in computational advances (Sequeiros-Borja et al., 2021). The integration of elegant algorithms and the expansion of software platforms have revolutionized our ability to analyze, predict, and engineer protein structures. Accessible web server platforms like VisualCMAT are one of many examples of computational tools that facilitate the design process, utilizing algorithms to provide an interface for researchers to explore enzyme design (Suplatov et al., 2018). Key to this analysis is the comprehensive exploration of enzyme structures, as structural information serves as a blueprint for rational design strategies.

How can predictive software be used to make inferences on how mutations affect residue structure and function, and how can this play a part in treating cancer and the future of medicine? Computational protein engineering offers a way for researchers to explore the inner workings of a protein that are difficult to observe in nature (Yakhini and Jurisica, 2011). For example, using information from protein interaction networks, researchers can map proteins onto larger networks and make connections through software. This was applied to APC, a tumor-suppressing gene that codes for tumor-suppressing proteins, that can fight colon cancer (Zheng et al., 2021).

Cancer, characterized by uncontrolled cell growth and proliferation, is a multifaceted disease with numerous genetic and molecular determinants (Gulati and Singh, 2024). Mutations in key enzymes play a pivotal role in the initiation and progression of cancer. Dysregulation of enzymes involved in critical cellular processes, such as DNA repair, cell cycle control, and apoptosis, can contribute to genomic instability and cancer traits (Gulati and Singh, 2024). For example, alterations in the activity of enzymes like kinases, which regulate signaling pathways controlling cell growth and survival, are commonly associated with cancer. Understanding the intricate interplay between enzymes, mutations, and cancer progression is crucial for the development of targeted therapies aimed at disrupting specific molecular pathways driving malignancy. For this study, I focused on enzymes isocitrate dehydrogenase (IDH1) and aconitase-2 (ACO2). These enzymes were chosen due to their drastic roles in the cell and their relation to cancer.

Isocitrate dehydrogenase 1 (IDH1), is an enzyme that has a regulatory role in the citric acid cycle (Murugan and Alzahrani, 2022). This metabolic pathway occurs in the matrix of the mitochondria that contributes to glycolysis and other organic molecules to feed into the oxidative phosphorylation (OXPHOS) pathway to generate cellular energy in the form of ATP. IDH1 catalyzes the conversion of isocitrate to alpha-ketoglutarate acid while also producing NADH (Murugan and Alzahrani, 2022). IDH1 relates to cancer in that mutations in the enzyme may lead to acute myeloid leukemia (AML) and/or gliomas (Murugan and Alzahrani, 2022). Furthermore, a mutation within the IDH1 gene can lead to a gain-of-function alteration of the enzyme's activity known as an R132H mutation. The harmful alteration of the enzyme would lead to an abnormal production of metabolic products, such as 2-hydroxyglutarate, which would disrupt cellular processes and cause uncontrollable growth.

The danger of IDH1 mutations can be seen in a study conducted by researchers to examine IDH1 mutation prevalence and link it as a significant factor in cancer. Within 40 cohorts of 14,726 cancers, IDH1 mutations were found in 3% of them (476) with the highest frequencies of IDH1 mutations resulting in oligodendrogliomas, anaplastic oligodendrogliomas, and diffuse astrocytomas (Murugan and Alzahrani, 2022). These three diseases being the most prevalent with an IDH1 is interesting as it demonstrates that IDH1 plays a role in an increased presence of cancer specifically within the brain as that is where these three cancers are located. While the percent number of IDH1 mutations found was seemingly minimal (3%), data collected suggest that the presence of the IDH1 enzyme is an indicator of better overall and progression-free survival (survival after cancer treatment) and is better overall for the prognosis of a patient going through cancer treatment (Murugan and Alzahrani, 2022). Patients with IDH1 mutations within gliomas are shown to have a median overall survival rate of 207 months and a free-progression survival rate of around 100 months. In contrast, patients with gliomas that don't have IDH1 mutations have a 25-month median overall survival rate and a 9-month free-progression survival rate. The implications of the research stipulate that specific IDH1 mutations play a significant role in brain tumors and other malignancies and that these mutations are beneficial to the prognosis of patients as they have a longer survival time before disease progression compared to cancers that do not have IDH1 mutations. Due to this, IDH1 mutations are not only a significant molecular identifier in gliomas, but also more beneficial to the prognosis of a patient due to a better survival rate (Murugan and Alzahrani, 2022).

Aconitase-2 (ACO2), similar to IDH1, also has a role in the citric acid cycle, catalyzing citric acid into isocitric acid (Ciccarone et al., 2020). The enzyme also has a role in suppressing tumors, maintaining genetic stability, and preventing mutated proteins from interfering with cell functions. Furthermore, changes in aconitase activity influence metabolic behavior in cancer cells, meaning that aconitase-2 is able to reprogram cancer cells to an extent. Using MCF-7 breast cancer cell line, a study measured ACO2 function in inhibiting cancer cell growth and tumorigenesis (Ciccarone et al., 2020). Using a multi-faceted approach, cells were treated and supplemented with various compounds (penicillin, chloroquine, BPTES, etc). Followed by analysis of protein and gene expression levels, as well degree of cell proliferation, ATP levels, and oxygen consumption. ACO2 was able to inhibit these cells by promoting oxidative metabolic activity and autophagic response. Overall, they found that ACO2 expression is reduced in breast cancer and that increasing its levels dampens MCF-7 cell proliferation, and that overexpression of the enzyme can induce autophagic/mitophagic flux in the instance of oxidative stress (Ciccarone et al., 2020).

Here, VisualCMAT, a web server designed to provide analysis on correlating mutations/co-evolving residues (two amino acids that show consistent similarities in evolution) was utilized to understand pivotal amino acid residues that exhibited engineering potential for therapeutics. The first step that took place was to create sequence information regarding the three enzymes. The alignment would consist of the same enzyme but the difference would be the type of species. Creating a sequence alignment between different species would convey the regions of similarities between the enzymes that could lead to functional and structural consequences as well as evolutionary relationships. Within the alignments of the enzymes, physical and structural properties of the sequences can be viewed, such as helix propensity, hydrophobic/hydrophilic, and more.

METHODS

Amino Acid Sequence Alignment

Amino acid sequences were extracted and aligned using the Uniprot database. At least four different amino acid sequences from mammalian species, except in the case of IDH1, were aligned for higher comparative analysis. For IDH1, the aligned species included; human, mouse, rat, *Ajellomyces capsulataus* (Darling's Disease - causing fungi species), and sheep. For caspase-2, the species that were aligned included; human, mouse, rat, and chicken. For aconitase-2, the species that were aligned included; human, mouse, rat, pig, and bovine (buffalo).

VisualCMAT analysis

To run VisualCMAT analysis, amino acid sequence alignments were created using the Uniprot database in the form of FASTA files. Experimental protein structures were collected from Protein Data Bank (PDB) including: IDH1 (PDB ID: 5k10), caspase-2 (PDB ID: 1PYO), and aconitase-2 (PDB ID: 1ACO). All protein structures derive from mammals, with 5K10 and 1ACO being in humans and 1PYO being in *Bos Taurus* (cattle). The corresponding PDB code, along with the alignment FASTA file, were run on the VisualCMAT program under the default settings.

Correlated Pair Analysis

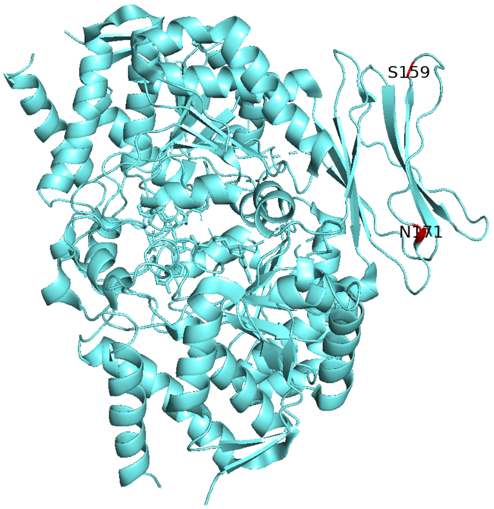
Structure and functional information were used to deduce and infer a rationale for specific mutation predictions that occurred on these enzyme's residues. The five lists for each protein are the predicted correlated pairs for the enzymes. The SEQ id's and the PDB id's are where the amino acids are placed on the protein for UniProt and the PDB respectively. UniProt provides information regarding the residue's location. To identify residues in general, I used the PDB ID since it refers to the 3D structure of the protein. Using the feature viewer, it can be determined where a residue is located, whether this is a beta strand, helix, binding site, or uncategorized. Furthermore, looking into an individual residue's specific function and structure would be useful in determining an interpretation for the residue pairs. For the two enzymes, UniProt's feature viewer accesses different information depending on the enzyme. For IDH1, the viewer is able to access binding sites, modified residues, helixes, beta strands, turns, and uncategorized residues. This is similar to aconitase-2 however information about secondary structures is not listed and instead replaced with information about its polar residues and disordered regions. UniProt does not have access to everything, however- if a residue is labeled as uncategorized, it often means ambiguity in determining the residue's specific classification.

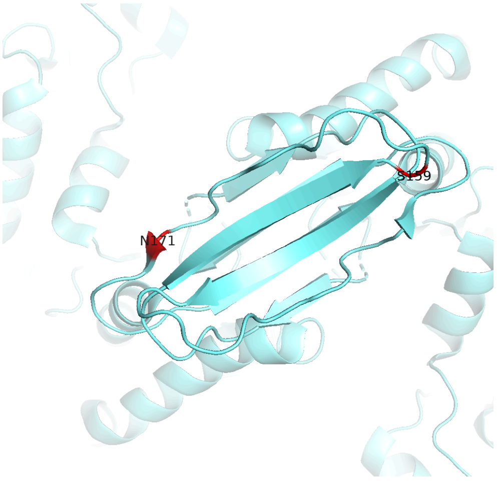
RESULTS
Table 1: IDH1 VisualCMAT:

LIST HEAD Rank	SEQ_id_i	PDB_id_i	SEQ_id_j	PDB_id_j	MIp	MIc	Zp	Zc
LIST DATA #1	S 151	///A/SER`159	N 163	///A/ASN`171	0.131	0.123	5.315	5.653
LIST DATA #2	K 1	///A/LYS`3	G 43	///A/GLY`45	0.127	0.113	5.135	5.233
LIST DATA #3	Q 155	///A/GLN`163	K 235	///A/LYS`243	0.113	0.106	4.592	4.942
LIST DATA #4	T 147	///A/THR`155	Q 155	///A/GLN`163	0.113	0.106	4.588	4.930
LIST DATA #5	T 147	///A/THR`155	K 235	///A/LYS`243	0.112	0.105	4.548	4.892

Dissecting the results of the VisualCMAT analysis, the "i" or the "j" in the id signifies whether it is the first amino acid in the pair or the second one. The datasets are listed in decreasing z-score, with list data #1 having the highest z-score and list data #5 having the lowest z-score. The z-score is significant because it shows how statistically significant the correlation between the two residues is. The higher the z-score, the higher the correlation between the two amino acids.

For IDH1, The top correlated pair is serine 159 and asparagine 171 (Table 1). I used the UniProt feature viewer to determine connections between the position of the residue with the overall protein. Serine 159 is part of the chain of the protein or the mature region of IDH1 as well as an antibody binding sequence that ranges from residues 144-208. A mature region is the region of the protein that follows post-translational modification, an overall process in which properties of a protein are changed either through breaking peptide bonds between residues within a protein and/or adding modifying groups. The antibody binding sequence is essentially segments of amino acids within the protein associated with binding interactions with antibodies, a protein synthesized by the immune system that binds with foreign/unwanted substances like germs and diseases and gets rid of them. I then used the GDC (Genomics Data Commons Data Portal) to see whether cancer mutations were near or directly on the residue locations. The 158 and 160 residues that are adjacent to the 159 residues have had missense mutations and these mutations had a moderate level of impact in the development of cancer, occurring in three cases. For the 171 residue, there is a missense mutation on the same three transcripts as the 159 residue.





The correlated pair in the second data set is lysine and glycine on residues 3 and 45, respectively (Table 1). Using Uniprot, both residues are not located on any beta strands or helixes, but they are both near modified residues: residue 2 and residue 42, Modified residues are amino acids that undergo a certain type of alteration, such as phosphorylation, acetylation, methylation, and more. Using the GDC, I analyzed whether there were cancer mutations near the residues or directly on them. Near residue 3, residue 5 has two types of mutations- a missense mutation with a moderate impact and a frameshift mutation with a high impact. There has only been one case of the missense mutation while the frameshift mutation has occurred twice. These mutations occurred in seven different cases. For residue 45, there is one case of a moderate impact missense mutation near it on residue 43. Like residue 3, this case appears on every transcript as well.

The correlated pair in the third data set is glutamine 163 and lysine 243 (Table 1). Both residues are not located on helixes, beta strands, or turns. They are both part of the chain/mature region of the enzyme. Glutamine 163 is also part of the 144-208 antibody binding sequence. On the GDC, residues 162 and 164, which are directly next to glutamine 163, have one case of missense mutations of moderate impact, occurring in three different cases. For lysine 243, There are no mutations that are directly on it or adjacent to the residue for all the transcripts.

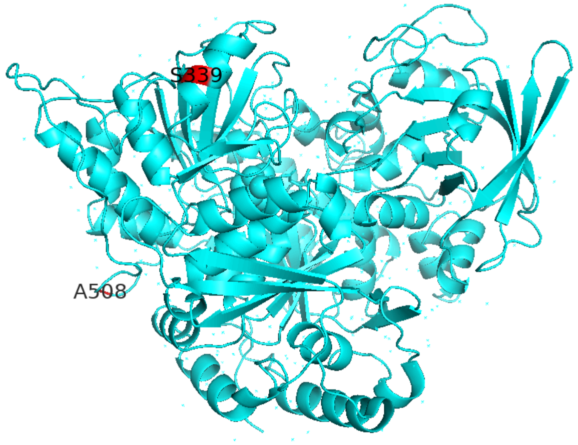
The correlated pair in the fourth data set is threonine 155 and glutamine 163 (Table 1). Threonine 155 is on a beta strand while glutamine is on neither a beta, helix, or turn. Both residues are part of the chain/mature region of the enzyme and are both part of the 144-208 antibody binding sequence. There has been one case of a missense mutation of moderate impact directly on the threonine residue. Residues 162 and 164, which are directly next to glutamine 163, have one case of missense mutations of moderate impact, present in three different cases.

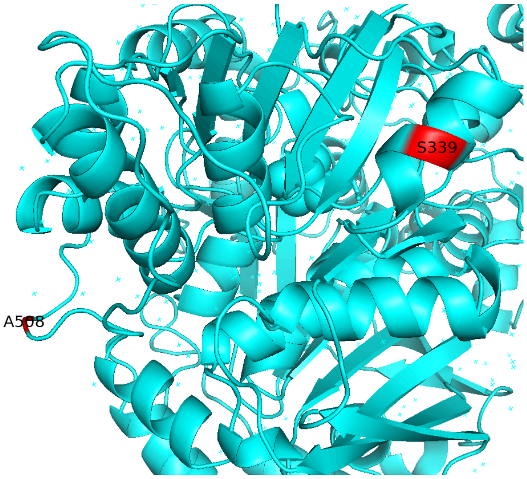
The correlated pair in the fifth data set is threonine 155 and lysine 243 (Table 1). Both amino acids are part of the chain/mature region of the protein. Threonine 155 is part of a beta strand as well as the 144-208 antibody binding sequence. Lysine 243 is however a modified residue, meaning it has some sort of modification akin to phosphorylation, acetylation, and other types. A missense mutation of moderate impact directly on the threonine 155 residue, occurring in three different cases (Table 1). For lysine 243, there are no recorded mutations on any of the transcripts.

Table 3: Aconitase-2 VisualCMAT:

LIST HEAD Rank	SEQ_id_i	PDB_id_i	SEQ_id_j	PDB_id_j	MIp	MIc	Zp	Zc
LIST DATA #1	S 336	///A/SER`339	A 505	///A/ALA`508	0.167	0.146	8.258	9.386
LIST DATA #2	S 336	///A/SER`339	R 525	///A/ARG`528	0.116	0.095	5.751	6.159
LIST DATA #3	A 505	///A/ALA`508	R 525	///A/ARG`528	0.114	0.093	5.636	6.040
LIST DATA #4	E 333	///A/GLU`336	E 506	///A/GLU`509	0.113	0.093	5.583	6.038
LIST DATA #5	E 333	///A/GLU`336	A 505	///A/ALA`508	0.113	0.093	5.571	6.016

For ACO2, the correlated pair in the first data set is serine 339 and alanine 508. Both residues are on the chain/mature region of the enzyme. Appearances of a single case of a missense mutation of moderate impact that occurs adjacent or directly on residues 339, over 12 mutations presented. For the second residue, alanine 508, a single case of a missense mutation of moderate impact occurs near or directly on the residue on residues 507 and 510 on, occurring in 7 different cases. Frameshift mutations of high impact on residue 499 are shown in two cases. The correlated pair in the second data set is serine 339 and arginine 528. Both residues are located on the chain/mature region of the protein. Arginine 528 is in a disordered region, meaning there is no secondary structure within that segment of the protein such as a beta sheet or alpha helix. It is also part of the 519-668 antibody binding sequence. There is a single case of a missense mutation of moderate impact on residue 526 on ENST00000396512 (805 aa), residue 525 on ENST00000676748 (747 aa), residue 529 on ENST00000216254 (780 aa), residue 524 on ENST00000676792 (725 aa), residues 523 and 525 on ENST00000677153 (747 aa), residue 526 on ENST00000678269 (805 aa), residue 527 on ENST00000677532 (788 aa), one case of a frameshift mutation of high impact on residue 532 on ENST00000677554 (716 aa), ENST00000679320 (731 aa), ENST00000216254 (780 aa), ENST00000678788 (775 aa), and none on ENST00000466237 (422 aa).





DISCUSSION

Computational biology and chemistry have made remarkable strides in the understanding of the biotic world. Scientists have developed models and utilized data analysis to comprehend the vast array of phenomena that occur within living organisms. The development of these models allows scientists to have a better understanding of biological and chemical processes that would be able to accelerate research in specific areas like combating diseases, such as cancer. Fundamentally, this technology has allowed for the deduction of protein folding predictions, different molecular interactions, and mutation predictions, all three aspects that are present in this paper.

In this paper, I analyzed the enzymes IDH1 and aconitase-2. These two enzymes have functions related to metabolic processes and mutations that affect these enzymes can be related to cancer. Mutagenesis of these enzymes correlates with an increased probability of DNA and mitochondrial damage that may result in malignant tumor growth. Through this paper, I used predictive software to infer correlations between mutations and the specific residues within these enzymes that can be affected by these mutations and thus lead to diseases like cancer. I utilized a protein alignment with different species and experimental protein structures from the Protein Data Bank (PDB) and inputted this into VisualCMAT, a software that can annotate specific predictions within co-evolving residue pairs. Through the data outputted by the software, I organized the pairs in the tables above, and the highest correlative numbers indicate a higher chance that these residues have co-evolved within the enzyme. Data can be gathered about these residue interactions through UniProt and the GDC data portal about their structural and functional purposes and certain mutations that have occurred on these residues. For example, within some correlated pairs, they reside in a region of a protein that functions in binding to proteins or their structure contains folds and structures like helices or beta sheets that affect their function. For example, an alpha helix sheet offers more environmental stability for a residue, allowing it to survive and function in more extreme temperatures/pHs. Furthermore, alpha helices allow the protein structure to be more compact and dense. A residue located on a beta-pleated sheet, on the other hand, is more flexible, flatter, and thinner. They form more flexible structures like fibrils and have more versatile formations. Through the GDC, previous cases of mutations on these enzymes can be viewed and the rate at which these mutations occur on the residues and the severity of the mutation provides valuable insight about the effect of mutations within these residues and how they can lead to cancer. It can then be inferred that these mutations that cause cancer have effects in damaging or changing the structure and function of the enzymes affected. For example, on a residue that has had a frameshift mutation, the mutation would have affected that residue's structural function within the enzyme, whether that be antibody binding or damage in their specific folding and structure. Through computational analysis, I was able to gather correlated pairs within IDH1 and aconitase-2 and analyze their structure and function, and infer based on mutations found through research how they might be able to affect residue pairs that are closely related to each other.

There is still some progress in making these conclusions more concrete. My research is only composed of analyzing the results of the predictive software but not how accurate they are. Protein analysis through UniProt, VisualCMAT, and pyMOL serves as a solid foundation for the computational analysis of enzymes, but there might be inconsistencies and inaccuracies in these predictions that need to be refined in the future. This could result in the incorrect residue pairs being correlated or a miscount in the amount of residues in an enzyme. To accurately test the capability of these predictive software, experiments should be run on these enzymes. These could take place in vivo experiments through cell culture or animal models. Then, inducing mutations within these select residues through directed-site mutagenesis and recombinant variant expression. Testing enzyme capability with these mutations in these specific residue-pairs such as collecting data on their thermostability, enzyme function, and structure would give a more clear picture of the capabilities of predictive software and how these mutations can affect these residues. Experimentation would thus give a more concrete picture of the effects these mutations have on the correlated pairs and the rate at which mutations occur can be a measure of how severe a mutation can be. With a better picture through experimentation, computational software can then be improved to be more accurate. Furthermore, through a more solid visual of these mutations' effects, researchers can then change the structures of these residues and proteins to be better equipped to fight against cancer. Biophysical research could also give more information on the structure and response.

These results show promise in not just protein engineering but also in a broad scientific scope. There are various applications for enzyme engineering and their effects in science, engineering, and medicine. In regards to biological science, furthering the capability of this predictive software can develop research on cancer and other diseases as well as create more preventive means of treatment through predictions. Bioorthogonal chemistry involves the use of bioorthogonal groups that “tag” biomolecules (Qin et al., 2018). Research and refinement in bioorthogonal chemistry can result in bioorthogonal molecules “tagging” cancer cells within the body (Qin et al., 2018). This would allow more targeted treatment of the disease with lower risks of side effects from radiation therapy and chemotherapy that result in hair loss and sickness as these treatments don’t have ways of differentiating healthy cells and cancer cells. Through bioorthogonal chemistry, differentiation would be possible, and combining more refined predictive software with this treatment would result in more accurate treatments of mutations and malignant growth in a patient, as predicting these mutations would be able to give medical professionals a clearer image of which biomolecules and cells to tag. While predictive software is very promising in the world of medicine and science, I believe it requires more testing and refinement to be fully actualized and reach its full potential as an efficient tool in preventing and targeting diseases.

ACKNOWLEDGMENTS

Thank you to Dr. Jose Delgado for his support, guidance, and mentorship throughout the research and writing processes of this paper.

REFERENCES

Ciccarone, F., Di Leo, L., Lazzarino, G., Maulucci, G., Di Giacinto, F., Tavazzi, B., & Ciriolo, M. R. (2020). Aconitase 2 inhibits the proliferation of MCF-7 cells promoting mitochondrial oxidative metabolism and ROS/FoxO1-mediated autophagic response. *British journal of cancer*, *122*(2), 182–193. https://doi.org/10.1038/s41416-019-0641-0

Gulati, P., & Singh, C. V. (2024). The Crucial Role of Molecular Biology in Cancer Therapy: A Comprehensive Review. *Cureus*, *16*(1), e52246. https://doi.org/10.7759/cureus.52246

Qin, L.-H., Hu, W., & Long, Y.-Q. (2018). Bioorthogonal Chemistry: Optimization and application updates during 2013–2017. *Tetrahedron Letters*, *59*(23), 2214–2228. https://doi.org/10.1016/j.tetlet.2018.04.058

Murugan, A. K., & Alzahrani, A. S. (2022). Isocitrate Dehydrogenase IDH1 and IDH2 Mutations in Human Cancer: Prognostic Implications for Gliomas. *British journal of biomedical science*, 79, 10208. https://doi.org/10.3389/bjbs.2021.10208

Sequeiros-Borja, C. E., Surpeta, B., & Brezovsky, J. (2021). Recent advances in user-friendly computational tools to engineer protein function. *Briefings in bioinformatics*, *22*(3), bbaa150. https://doi.org/10.1093/bib/bbaa150

Suplatov, D., Sharapova, Y., Timonina, D., Kopylov, K., & Švedas, V. (2018). The visualCMAT: A web-server to select and interpret correlated mutations/co-evolving residues in protein families. *Journal of bioinformatics and computational biology*, *16*(2), 1840005. https://doi.org/10.1142/S021972001840005X

Yakhini, Z., & Jurisica, I. (2011). Cancer computational biology. *BMC bioinformatics*, *12*, 120. https://doi.org/10.1186/1471-2105-12-120

Zheng, L., Zhan, Y., Lu, J., Hu, J., & Kong, D. (2021). A prognostic predictive model constituted with gene mutations of *APC*, *BRCA2*, *CDH1*, *SMO*, and *TSC2* in colorectal cancer. *Annals of translational medicine*, *9*(8), 680. https://doi.org/10.21037/atm-21-1010