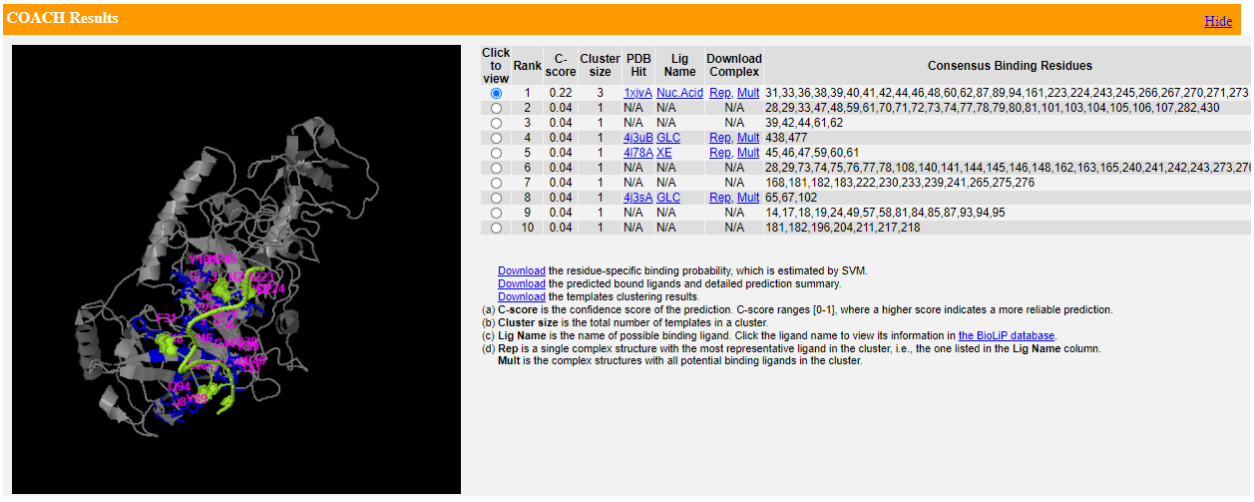
Prediction Software

Many bioinformatic tools involve the use of algorithms to estimate genomic structures, functions, or attributes based solely on genetic, proteomic sequences. Throughout my time in the JHU Individualized Genomics & Health program, these programs have allowed for insight into otherwise meaningless data. Presented in this document will be examples mainly from the Bioinformatics: Tools for Genome Analysis course, as much of the work done in this course was accomplished with unannotated/new sequence reads in which these prediction programs helped tremendously for analysis. Specifically, an analysis of predicted protein stability changes was a necessary step in the analysis of potential functional changes caused by the rs75932146 SNP. Also, the relation that POT1 had to other genes within its network was examined for potential influences that could also affect patients with this SNP and cancer. The information presented below is my own research on the POT1 gene's associations and structural changes found due to the V326A mutation, with an in depth analysis presented in the research paper.

COACH: Protein to ligand binding site prediction

COACH is a server-side protein to ligand binding site prediction software that is hosted by the University of Michigan[1,2]. After protein structure prediction is completed via submission of the protein's PDB or FASTA file, COACH will utilize TM-SITE and S-SITE prediction to analyze and present possible binding sites within the protein[1,2]. COACH also incorporates I-TASSER(see the above) to generate it's 3D models[1,2]. We used COACH to analyze structural changes and binding site changes that occurred when comparing the results from the normal POT1 protein sequence and with the alanine change at position 326 in the POT1 protein with SNP rs75932146.



COACH Results

Hide

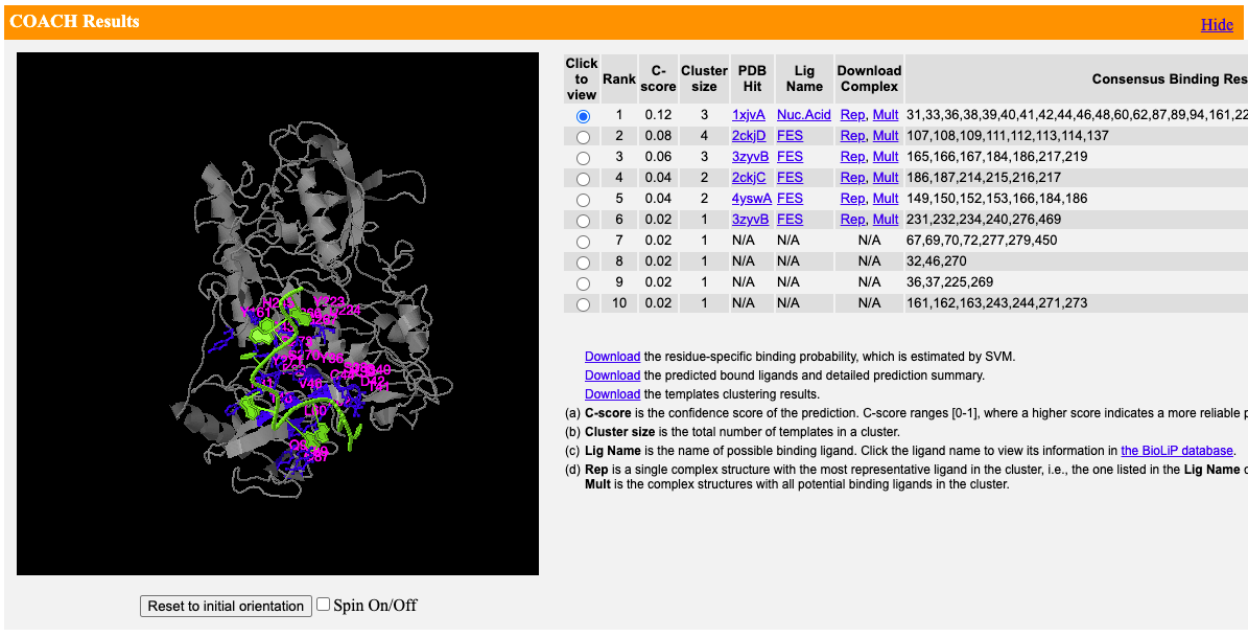
Click to view	Rank	C-score	Cluster size	PDB Hit	Lig Name	Download Complex	Consensus Binding Residues
◉	1	0.22	3	1xjvA	Nuc.Acid	Rep, Mult	31,33,36,38,39,40,41,42,44,46,48,60,62,87,89,94,161,223,224,243,245,266,267,270,271,273
○	2	0.04	1	N/A	N/A	N/A	28,29,33,47,48,59,61,70,71,72,73,74,77,78,79,80,81,101,103,104,105,106,107,282,430
○	3	0.04	1	N/A	N/A	N/A	39,42,44,61,62
○	4	0.04	1	4j3uB	GLC	Rep, Mult	438,477
○	5	0.04	1	4l78A	XE	Rep, Mult	45,46,47,59,60,61
○	6	0.04	1	N/A	N/A	N/A	28,29,73,74,75,76,77,78,108,140,141,144,145,146,148,162,163,165,240,241,242,243,273,27(
○	7	0.04	1	N/A	N/A	N/A	168,181,182,183,222,230,233,239,241,265,275,276
○	8	0.04	1	4j3sA	GLC	Rep, Mult	65,67,102
○	9	0.04	1	N/A	N/A	N/A	14,17,18,19,24,49,57,58,81,84,85,87,93,94,95
○	10	0.04	1	N/A	N/A	N/A	181,182,196,204,211,217,218

Download the residue-specific binding probability, which is estimated by SVM.
Download the predicted bound ligands and detailed prediction summary.
Download the templates clustering results.
(a) **C-score** is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable prediction.
(b) **Cluster size** is the total number of templates in a cluster.
(c) **Lig Name** is the name of possible binding ligand. Click the ligand name to view its information in the BioLiP database.
(d) **Rep** is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the **Lig Name** column.
Mult is the complex structures with all potential binding ligands in the cluster.

- Normal POT1 protein results from COACH. Predicted ligand(in this case, Nucleic Acid) binding site at the bottom of the model.

The prediction of the Normal POT1 protein showed a confidence score of 0.22 in the predicted nucleic acid binding site. Along with this, the protein showed a confidence score of 0.40 for the binding-specific

substrate comparison site (TM-SITE), and a score of 0.32 for the sequence profile alignment(S-SITE) prediction.



COACH Results Hide

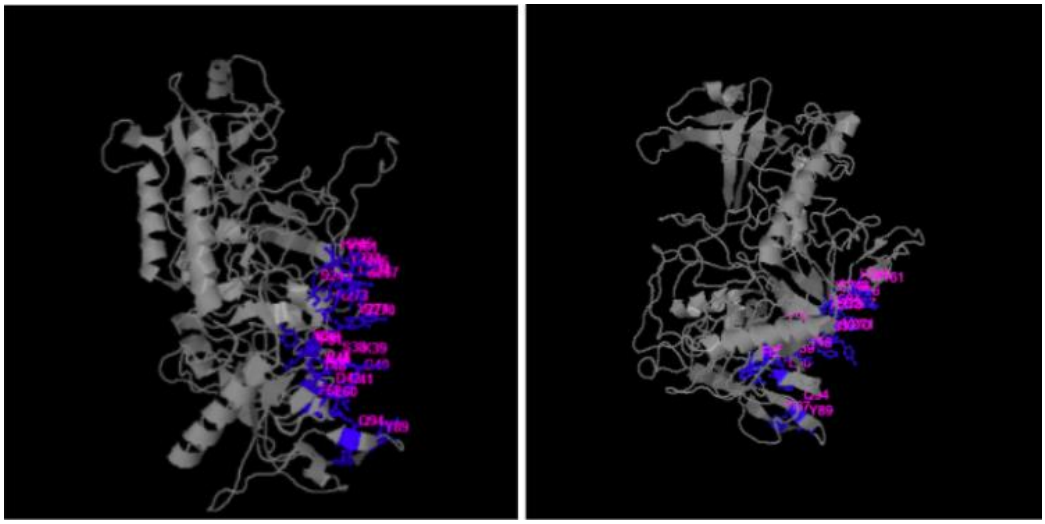
Click to view	Rank	C-score	Cluster size	PDB Hit	Lig Name	Download Complex	Consensus Binding Res
◉	1	0.12	3	1xjvA	Nuc.Acid	Rep, Mult	31,33,36,38,39,40,41,42,44,46,48,60,62,87,89,94,161,22
○	2	0.08	4	2ckjD	FES	Rep, Mult	107,108,109,111,112,113,114,137
○	3	0.06	3	3zyvB	FES	Rep, Mult	165,166,167,184,186,217,219
○	4	0.04	2	2ckjC	FES	Rep, Mult	186,187,214,215,216,217
○	5	0.04	2	4yswA	FES	Rep, Mult	149,150,152,153,166,184,186
○	6	0.02	1	3zyvB	FES	Rep, Mult	231,232,234,240,276,469
○	7	0.02	1	N/A	N/A	N/A	67,69,70,72,277,279,450
○	8	0.02	1	N/A	N/A	N/A	32,46,270
○	9	0.02	1	N/A	N/A	N/A	36,37,225,269
○	10	0.02	1	N/A	N/A	N/A	161,162,163,243,244,271,273

Download the residue-specific binding probability, which is estimated by SVM.
Download the predicted bound ligands and detailed prediction summary.
Download the templates clustering results.
(a) **C-score** is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable p
(b) **Cluster size** is the total number of templates in a cluster.
(c) **Lig Name** is the name of possible binding ligand. Click the ligand name to view its information in the BioLiP database.
(d) **Rep** is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the **Lig Name** c
Mult is the complex structures with all potential binding ligands in the cluster.

Reset to initial orientation ☐ Spin On/Off

- Mutant POT1 protein results from COACH. Predicted ligand binding site is seen once again at the bottom of the model.

Comparatively, there is a change in not only confidence scores when the V326A change is predicted for using COACH, but also a slight visual change to the overall structure of the protein when comparing the two different predictions(image shown below). The nucleic acid binding site now has a confidence score of 0.12, lower than the original and normal proteins confidence. The TM-SITE predicted structure has a confidence score of 0.39 and the S-SITE score is 0.32. These differences can be attributed to a number of components related to how the COACH program handles using similar protein sequences, but must also be assumed to have some part to do with the V326A change in the mutant POT1 protein.



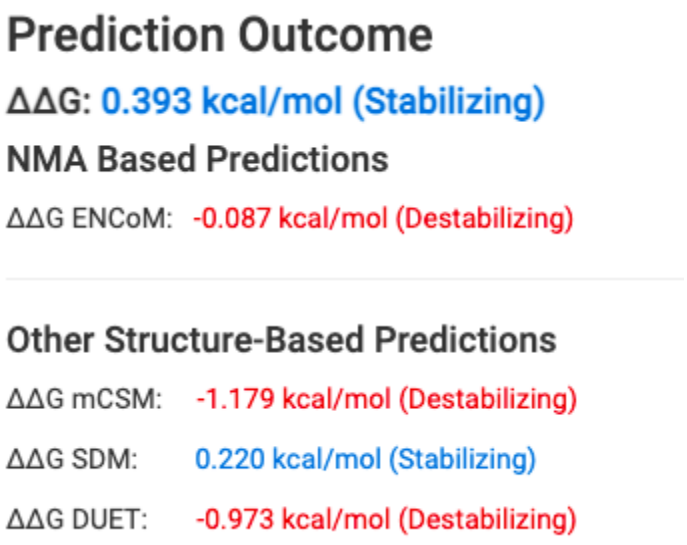
- A comparison of the visual predicted protein structure between the normal POT1 protein(left), and the mutant POT1 protein(right) from COACH.

Multiple prediction tools were used to analyze the predicted functional effect of the V326A change found in the rs75932146 mutant POT1 protein.

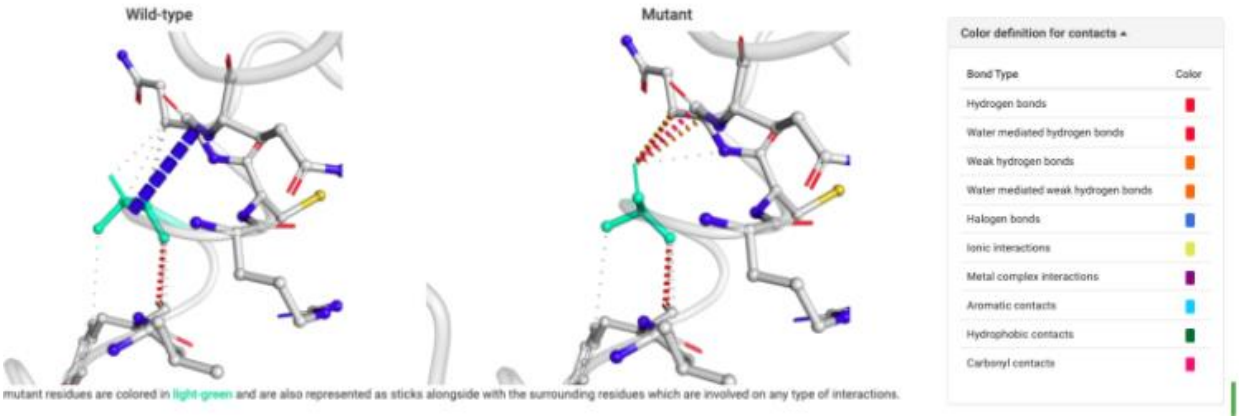
Prediction Tool	Predicted Confidence	Predicted Outcome
SIFT[3]	0.53	Tolerated
PolyPhen-2[4]	0.084	Benign
SDM[5]	0.12	Increased Stability
PredictSNP[6]	0.83	Neutral
MAPP[6]	0.68	Benign
PhD-SNP[6]	0.83	Neutral
SNAP[6]	0.55	Neutral
PANTHER[6]	0.57	Neutral

Table 3 demonstrates multiple tools utilized in the prediction of functional changes that are associated with the mutation of the POT1 protein. PredictSNP[6] is a program capable of collecting information from and executing predictions from multiple other tools, including MAPP, PhD-SNP, SNAP, and PANTHER, to name a few. From table 1, it becomes clear that although many changes to the structure of the protein can be seen with the mutation of the POT1 gene, many prediction tools show that these changes do not interfere with the overall function of the protein. The position of 326 is not seen near the protein-ligand site and does not represent a significant switch in terms of amino acid composition, and these results are indicative of this.

The program DynaMut[7] was also used to determine protein structure stability, and showed a different result to those shown in table 3.



- **Figure 24**: DynaMut results demonstrating protein stability differences introduced by the POT1 gene mutation. Along with the DynaMut prediction (stabilizing), the software also included results from ENCoM, mCSMSDM(which matches results from the above table) and DUET.



- **Figure 25**: DynaMut also provides a prediction of differences of atomic interactions introduced by protein mutations. In the wild type, a halogen bond can bee seen forming from the 326 amino acid position(shown in light green) to a neighboring amino acid. This changes to hydrogen bonds with the mutation from valine to alanine.

Figure 25 shows an interesting juxtaposition from the results seen in table 3, and agrees with the mCSM results shown in figure 23. Therefore, it cannot be concluded confidently that the V326A amino acid change does not have a destabilizing effect on the protein, as many results such as different structural changes and interatomic interactions demonstrated in this paper have shown that this SNP has some impact on the protein.

References:

1. Jianyi Yang, Ambrish Roy, and Yang Zhang. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment, Bioinformatics, 29:2588-2595 (2013).
2. Jianyi Yang, Ambrish Roy, and Yang Zhang. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions, Nucleic Acids Research, 41: D1096-D1103 (2013).
3. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001 May;11(5):863-74. doi: 10.1101/gr.176601. PMID: 11337480; PMCID: PMC311071.
4. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature methods, 7(4), 248–249. https://doi.org/10.1038/nmeth0410-248
5. Arun Prasad Pandurangan, Bernardo Ochoa-Montaño, David B. Ascher, Tom L. Blundell,. (2017). SDM: a server for predicting effects of mutations on protein stability. Nucleic Acids Research. Volume 45, Issue W1, Pages W229–W235, https://doi.org/10.1093/nar/gkx439
6. Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J., Damborsky, J., 2014: PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. PLOS Computational Biology 10: e1003440.
7. Carlos HM Rodrigues, Douglas EV Pires, David B Ascher. (2018). DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Research. Volume 46, Issue W1, Pages W350–W355, https://doi.org/10.1093/nar/gky300