Predicting Pharmacogenetic Responses to Prescription Drugs via Protein Modeling
Warfarin and CYP2C9

Aiyana Parker
12/10/2022

Video Tutorial: Robetta Tutorial.mp4

Purpose: Investigate structural and functional effects of mutations in genes known to have roles in drug metabolism.

Introduction

It is known that many individuals have variable responses and side effects to prescription drugs due to differences in their genetic makeup. Relevant variants are often found in enzymes responsible for the metabolism of the drug in question or the metabolism of an unrelated drug, nutrient, etc., that reacts with the drug in question; though effects on initial drug uptake or the disease pathway associated with this drug have been implicated as well (Roden et al., 2011). Some of these variants may cause adverse but mostly benign side effects such as headaches or nausea, and some of them may render a certain drug ineffective leading the prescriber to switch the method of treatment; but some variants can cause serious consequences if they are undetected before administration of the drug (Meyer, 2000). One example of a more complication-inclined variant drug response is that of the anticoagulant warfarin.

Various mutations in the enzymes associated with warfarin metabolism have been found to cause either warfarin resistance or accumulation, which can lead to thromboembolism or bleeding disorders, respectively, if dosing is not adjusted accordingly (Wiedermann & Stockner, 2008). Several genes, primarily CYP2C9 and VKORC1, have known variants implicated in these variable pharmacogenetic responses, and trends can be seen within populations of shared ancestry. There is still a relatively high amount of individual variation even within these populations, however, leading to warfarin dosing being a very sensitive process. Intensive monitoring of clotting factor levels and very delicate dose tapering are two methods used to try

to prevent adverse events (Johnson et al., 2017). In order to avoid long periods of experimental dosing and the possibility of fatal complications, it would be beneficial to be able to predict these responses before they occur. Bioinformatics offers a possible solution to this problem in the form of predictive models and algorithms that can determine the effect of mutations on the structure of proteins as well as the functions they are involved in.

One such predictive algorithm is the protein modeling algorithm Robetta (https://robetta.bakerlab.org/), which will be used throughout this manual and the accompanying tutorial. Robetta is a protein structure prediction tool based primarily on the deep learning algorithm RoseTTAFold developed and run by the Baker Lab at the University of Washington. There are other methods available within Robetta for comparative modeling and protein domain predictions, but RoseTTAFold is the most accurate and comprehensive of the options. RoseTTAFold takes an input of a protein sequence between 27-1200 amino acids in length and returns a set of 5 predictive models along with statistical measures of confidence.

The input used for this project was the 490aa protein FASTA sequence of the cytochrome P450 2C9 precursor peptide (NP_000762.2). The wild-type sequence as well as an assortment of manually entered variants were submitted to the RoseTTAFold queue for predictive modeling and visual comparison. Variants modeled were found through a combination of literature searches as well as the dbSNP database.

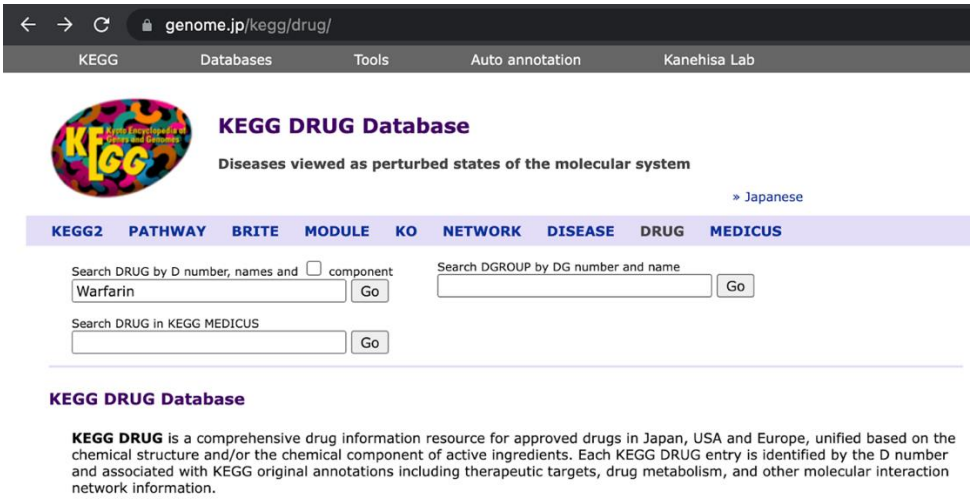
Manual

Step 1: Determine Genes of Interest

The first step to investigating pharmacogenetic drug responses is to identify relevant genes of interest. These could be genes with:

- Known clinical significance
- Role in drug metabolism
- Role in target disease pathways

A google scholar or PubMed search of “[drug] pharmacogenetics” or “[drug] metabolic pathway” could be a good place to start, or you could utilize a database such as KEGG (https://www.genome.jp/kegg/drug/) which will offer insight into therapeutic targets, metabolism, and other associated pathways.



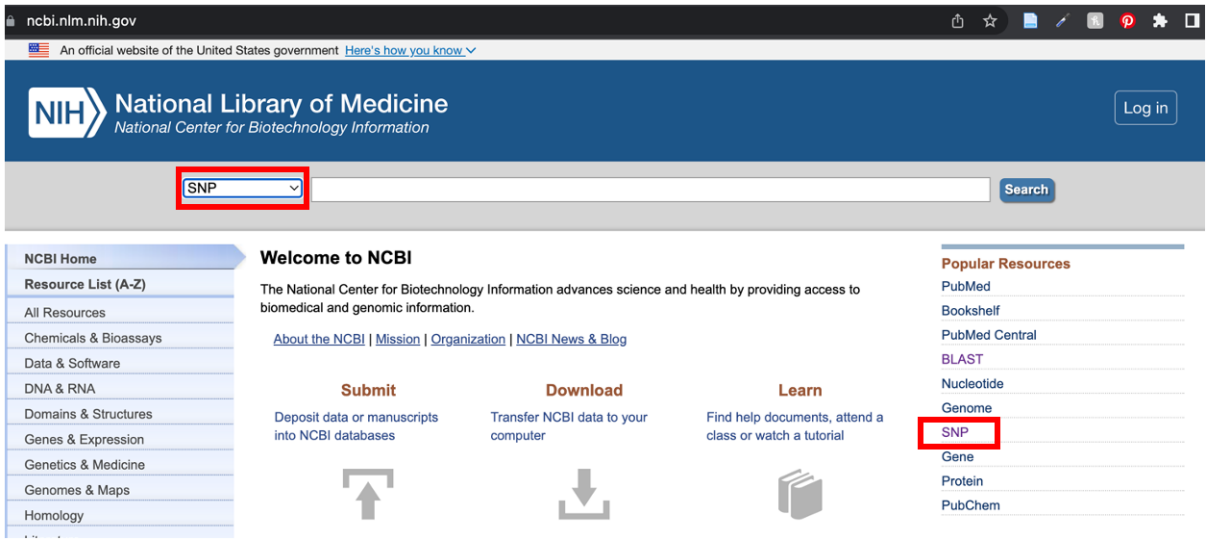
I recommend creating a table to keep track of genes and relevant information found during this stage. Below is an example using warfarin, an example which will be continued throughout this manual and the accompanying video tutorial.

Gene	Relevant Info	Sources
CYP2C9	Cytochrome P450 isoform involved in oxidation Primary metabolizer of more potent enantiomer S-warfarin (and many other drugs) Most common variants associated with lower dose requirements	https://ascpt.onlinelibrary.wiley.com/doi/10.1002/cpt.668 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3201766/
VKORC1	Vitamin K reductase → inhibited by warfarin Associated with clotting factor deficiency Associated with warfarin resistance Mostly promoter variants	https://www.kegg.jp/entry/hsa:79001 https://ascpt.onlinelibrary.wiley.com/doi/10.1002/cpt.668 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3093198/
NQO1	Catalyzes vitamin K reduction → promotes clotting Inconclusive findings on importance, may be backup pathways to compensate	https://www.kegg.jp/entry/hsa:1728 https://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-018-0837-x#:~:text=NQO1%20is%20thought%20to%20reduce,that%20can%20affect%20warfarin%20dose
…		

Step 2: Identify Mutations

Once you have identified your gene(s) of interest, the next step is to identify variants in those genes you wish to model and investigate. Your initial literature search may yield some results, but additional inquiry will likely be needed. I utilized dbSNP for this stage, and I will illustrate how below.

1. First, navigate to the dbSNP homepage either through this link (https://www.ncbi.nlm.nih.gov/snp/) or via the NCBI homepage. Select *SNP* in the dropdown menu next to the search bar or click the **SNP** link under the *Popular Resources* header on the right-hand side of the page.



2. Type your gene symbol into the search bar at the top and hit enter

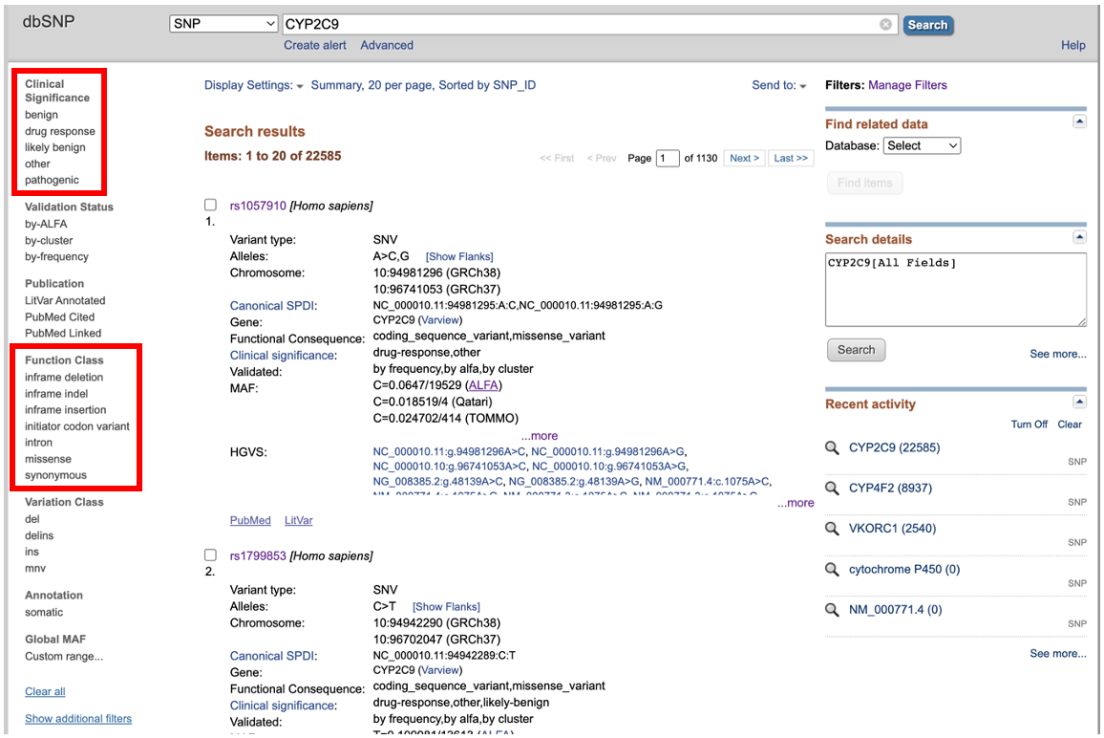
   Note: If it brings you back to the search homepage, this means there were no results in dbSNP and you need to alter your search term. e.g., CYP2C9 works but the full name cytochrome P450 family 2 subfamily C member 9 does not

   Note 2: The first column of this table should work every time

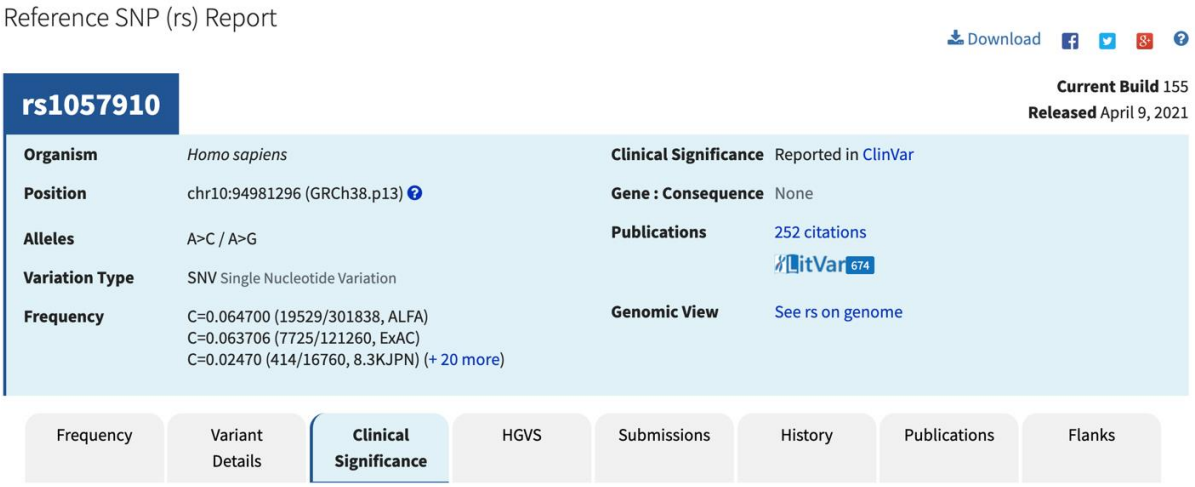
   https://tools.thermofisher.com/content/sfs/brochures/cms_042013.pdf

3. You will likely get thousands of results unless you are dealing with an uncommon gene. Use the filters on the left-hand side to narrow these results down and make them more

manageable. I utilized “drug response” under **Clinical Significance** and “missense” under **Function Class** to bring my results from 22585 to 16.



4. Clicking on any of the entries will bring you to a separate page that lists all of the information found in the search list as well as extensive additional information, including allele frequencies in different populations, details about the clinical significance of the variant, and links to PubMed publications discussing this variant.



5. Create another table to keep track of relevant mutations found, including nucleotide and protein changes and any variant IDs that may be useful.

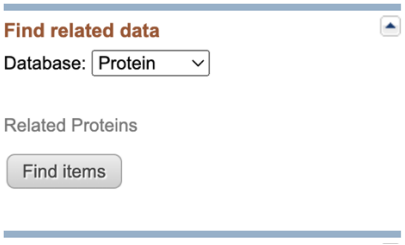
   Note: No variant IDs are needed for the tools used in this manual, but many other databases or predictive algorithms such as VEP and Poly-Phen do utilize them. rsIDs are also used as entry identifiers in dbSNP.

Gene	Nucleotide Change	Protein Change	Variant ID
CYP2C9	1075A>C,G	Ile359Leu,Val	rs1057910
CYP2C9	430C>T	Arg144Cys	rs1799853
CYP2C9	449G>A,C,T	Arg150His,Pro,Leu	rs7900194
CYP2C9	1003C>T	Arg335Trp	rs28371685
CYP2C9	1080C>A,G	Asp360Glu	rs28371686
VKORC1	-1639G>A	N/A	rs9923231
VKORC1	106G>A,T	Asp36Asn,Tyr	rs61742245
VKORC1	383T>G	Leu128Arg	rs104894542
…			

Step 3: Access Appropriate Dataset(s)

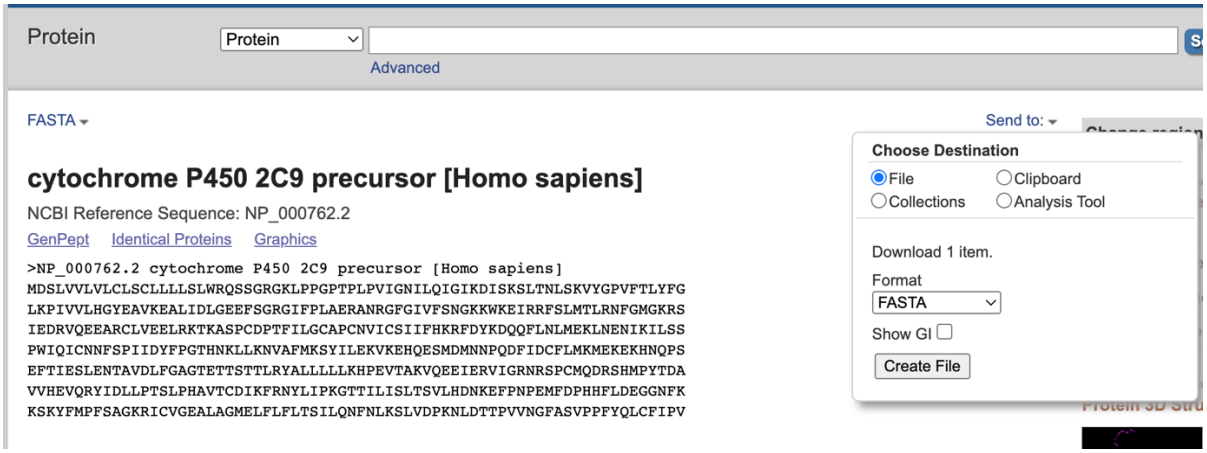
Now that you have determined the mutations found in your gene(s)/protein(s), you will need the wild-type protein sequence to use as the base of your modeling.

1. Return to the main Search results page. On the right-hand side under **Find related data**, select the *Protein* database and click **Find items**.



   Note: You can also search the protein database the same way we searched the SNP database in the first step of section 2, but you will likely have to filter through protein sequences of varying completeness.

2. This will bring you to search results for the protein in which these variants were identified. There may be multiple results, choose one of them and click on its title. This page will have a summary of information and publications about this protein which you can read through if you wish.
3. Under the title at the top of the page, click the button that says FASTA, which will bring you to a page with just the protein sequence which is what we need. You will be able to copy this sequence directly into the modeling program or you can download the FASTA file from this page for future reference.
4. Under the **Send to:** drop down, select "File" under **Choose Destination**, make sure the **Format** box says "FASTA", and then click **Create File**. This will download a file called sequence.fasta to your computer which you can rename and store wherever you wish.



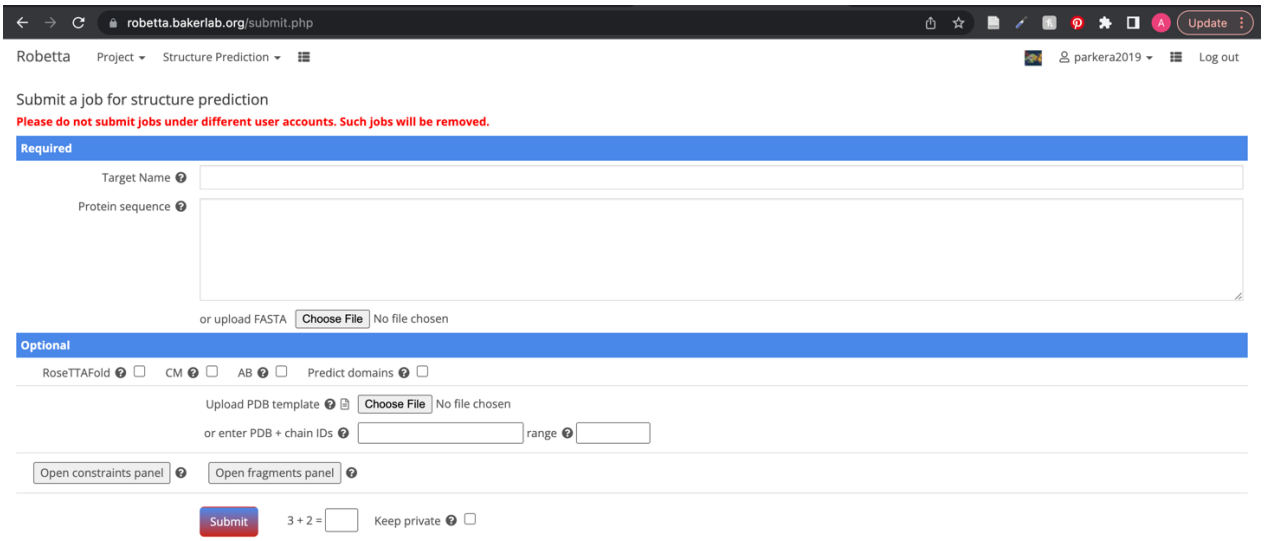
5. Repeat for all of the protein sequences you will be modeling. For the purposes of this tutorial, I used only the cytochrome P450 2C9 precursor protein to limit the complexity of the example.

Step 4: Query Submission

Once you have obtained the wild-type protein sequence for your gene of interest and a list of variants you wish to model, you are ready to begin modeling.

1. Navigate to Robetta using this link https://robetta.bakerlab.org/ or by searching "robetta" in your search engine. It should be the top link.

2. Before you begin using Robetta, you will have to register for a free account. Click the **Register** link in the top right of the home page, input all of the required information, then click **Register** at the bottom. Once you registered and confirmed your account, you can begin modeling.
3. Select **Submit** under **Structure Prediction** at the top of the home page. It will bring you to a page that looks like this:

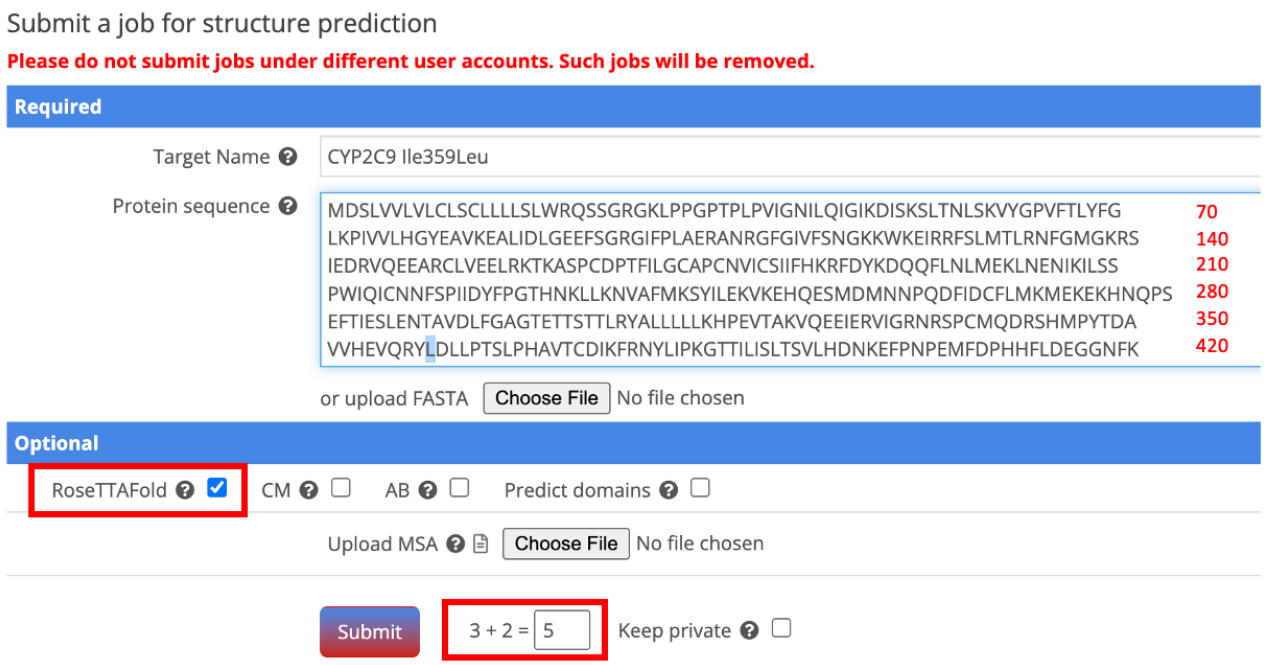


4. Choose a name for your job that you will be able to identify and type that in the **Target Name Box**, then paste your saved protein sequence (without the FASTA header) in the **Protein Sequence** box.
5. If you would like to model the wild-type protein at this point, skip to step 6. Otherwise, manually insert the mutation(s) into the protein sequence. If you copy and paste the sequence directly from the protein FASTA page, there are exactly 70 amino acids per line; and this formatting is conserved by the Robetta submission box, which simplifies this process. One benefit of this modeling program is that it allows for the entry of as many mutations as you would like since you can edit the sequence yourself and it is not based on variant IDs. I used a mixture of single (Ile359Leu) and multiple adjacent (Arg335Trp + Ile359Leu) SNPs for the models in this project.

Note: There are public applications to insert mutations for you if you wish to use those. Some of them may require variant IDs or the nucleotide sequence, which can be found on NCBI.

Note 2: You can also upload a FASTA file rather than pasting the sequence, though if you choose to manually insert mutations, remember to edit the FASTA file before uploading.

6. Check the box next to **RoseTTAFold** to make sure you are using the correct algorithm, answer the math problem verification, and click **Submit**.



7. Repeat this process for all the mutations you wish to model. The algorithm does not provide instantaneous results and may take up to several days to run. You should receive an email when each of your models is complete with a link to the query results page for your submission.

Step 5: Analysis

1. If the link provided in your confirmation email does not work, or if you would like to check the status of your query before the email is received, click on **Queue** under **Structure**

Prediction. This will bring you to a page of all public queries with a small amount of information on each one. All public queries are accessible to all users, but they do expire after a little over a month, so there is no historical record created.

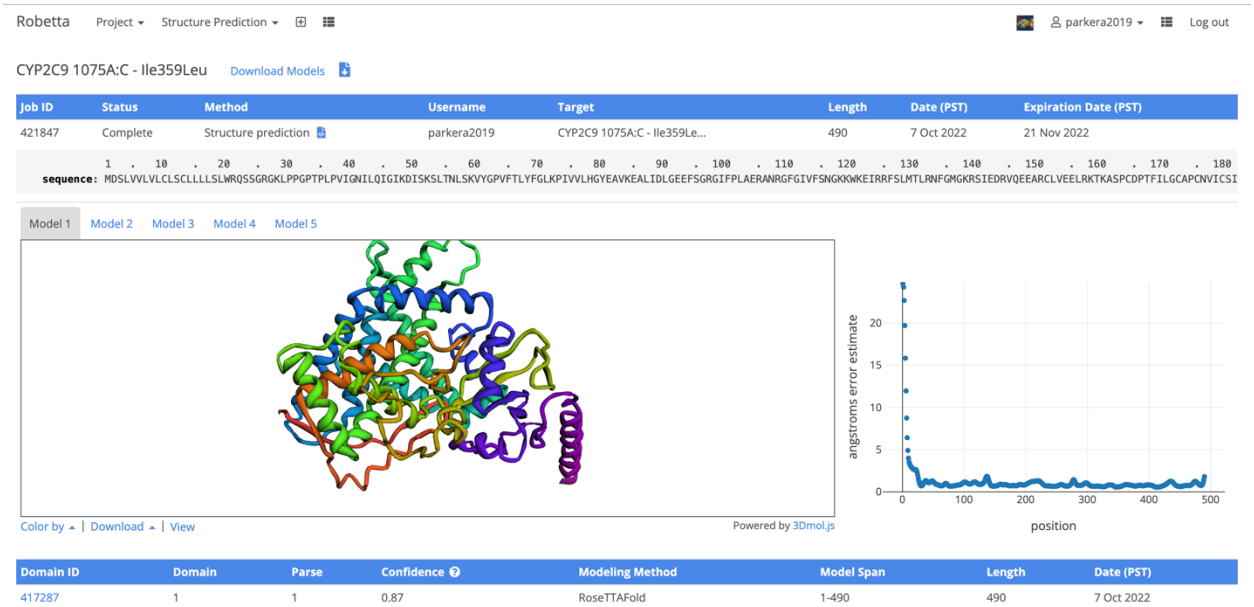
2. I recommend filtering by your username, though you can also filter by the Target Name you input during submission. This will bring you to a page that looks like the one seen below. A status of **Complete** means that your structure predictions are ready, and a status of **Active** means they are still being worked on.

Structure Prediction Queue

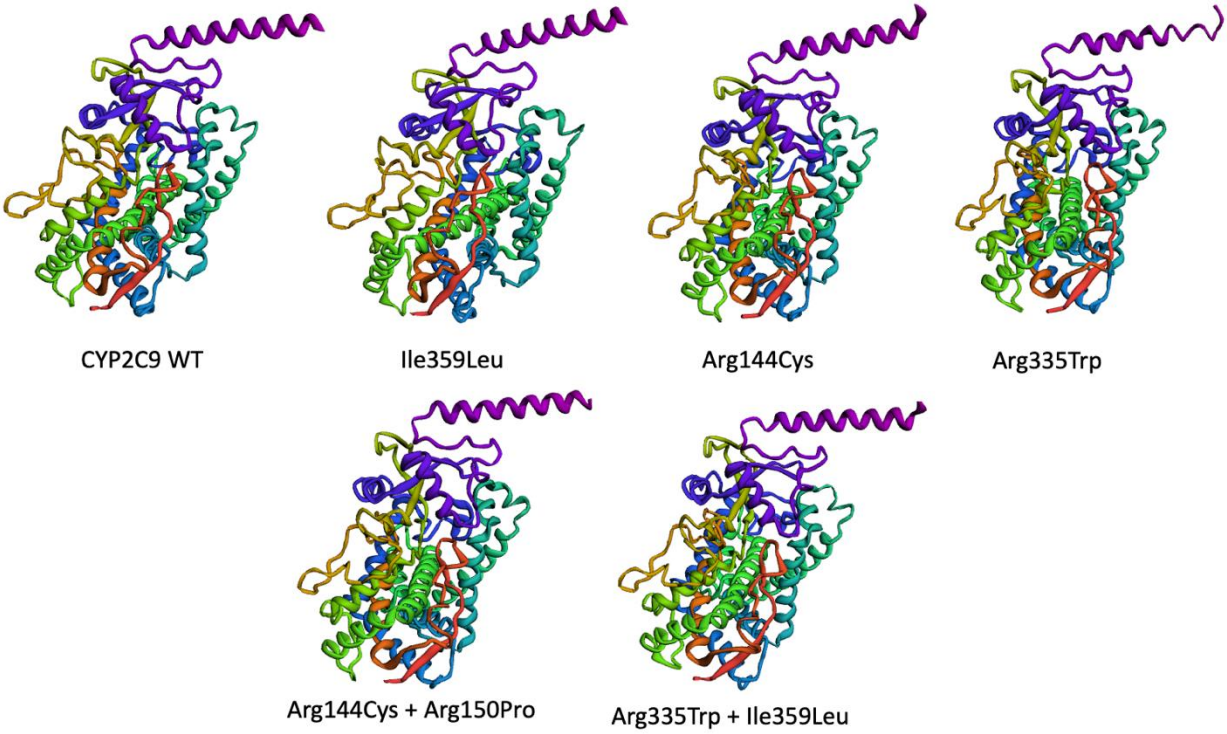
ID: Username: parkera2019 Target: Sequence: Search

Job ID	Status	Method	Username	Target	Length	Date (PST)	Expiration Date (PST)
435084	Complete	Structure prediction	parkera2019	CYP2C9 Precursor: Arg144C...	490	22 Oct 2022	7 Dec 2022
435083	Complete	Structure prediction	parkera2019	CYP2C9: Arg335Trp and Ile...	490	22 Oct 2022	7 Dec 2022
435082	Complete	Structure prediction	parkera2019	CYP2C9 Precursor: Arg150P...	490	22 Oct 2022	7 Dec 2022
435081	Complete	Structure prediction	parkera2019	CYP2C9 Precursor: Arg335T...	490	22 Oct 2022	7 Dec 2022
435080	Complete	Structure prediction	parkera2019	CYP2C9 Precursor: Arg150H...	490	22 Oct 2022	7 Dec 2022
435079	Complete	Structure prediction	parkera2019	CYP2C9 Precursor: Arg144C...	490	22 Oct 2022	7 Dec 2022
427061	Complete	Structure prediction	parkera2019	CYP2C9	490	12 Oct 2022	28 Nov 2022
421847	Complete	Structure prediction	parkera2019	CYP2C9 1075A:C - Ile359Le...	490	7 Oct 2022	21 Nov 2022

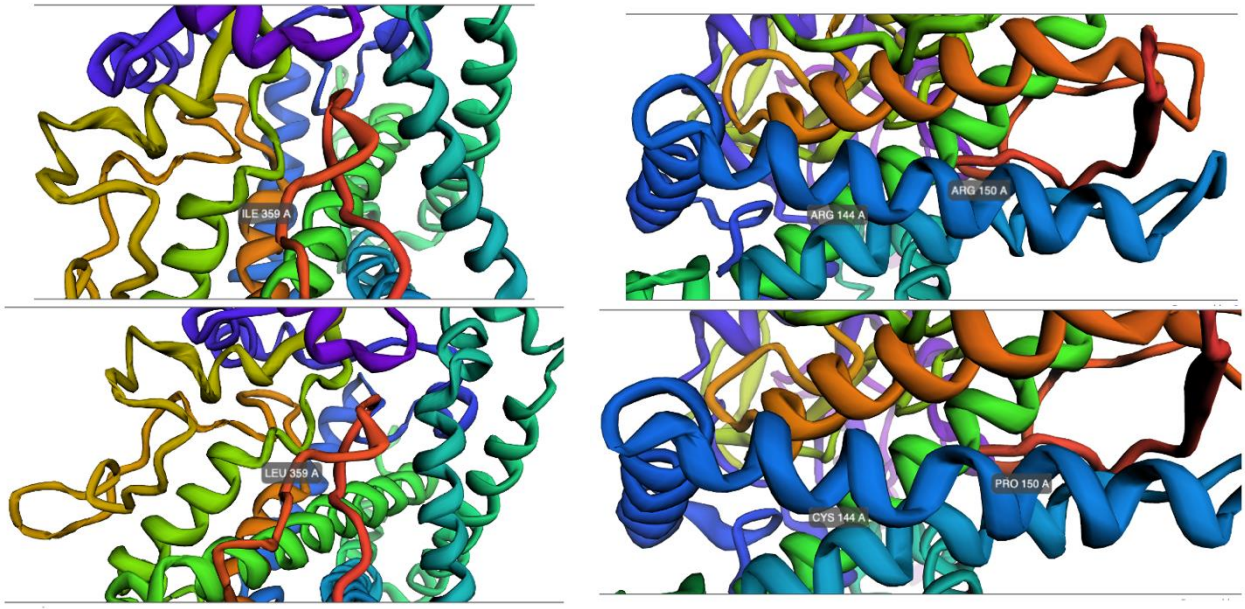
3. Click the link under Job ID to be brought to the results page for any of your submissions. This page shows you the sequence you input at the top, along with the top 5 predicted models and a graph of how confident the algorithm is about the positioning of each amino acid for each model. These models can be downloaded as a .pdb file with the position of every atom in the model. This can be viewed in a text editor, though very little information can be gleaned from this file without another modeling program. At the bottom of the page, RoseTTAFold also provides a GDT (Global Distance Test) confidence level which is a measure of positional agreement between the 5 predictions it provided.



4. Compare the overall shape of your variant models and the wild-type protein and the different protein domains within them to see if your mutations had any drastic effect on the protein structure. As you can see below, this was not the case for any of the CYP2C9 variants I modeled. All the original domains are consistently present, and I did not observe any major changes in folding. There are some slight positional variations, such as the angle at which the lime green domain is situated or the distance of the yellow domain from the rest of the body of the protein. Unfortunately this program does not allow for direct comparison of two models, so it is difficult to tell if these changes are legitimate or due to slight differences in viewing angle. If truly present, these slight variations may be enough to disrupt warfarin binding and cause the clinical effects observed; but investigation of the structure at a finer scale may also grant more insight.



5. You can switch between the 5 models, as well as adjust the viewing angle and zoom in on any of them. If you hover over the model, it will tell you the amino acid and its position within the sequence. Use this to find the spot in the protein that your mutation(s) are within and compare the structure immediately flanking this position across your variant models to see if the mutation altered the structure at a finer scale. Closely examine the structure of the folds directly involved with the mutation, as well as the interactions between that area and other nearby domains. Once again, I did not observe any major changes in the structures of the mutated proteins compared to wildtype. There are some slight striations and shadows in the folds that may indicate structural deviations, but the possible impact of these cannot be evaluated without further research into the function of the individual domains in the protein. Two examples are shown below of the Ile359Leu mutation as well as the combo of Arg144Cys and Arg150Pro.



6. Continue to play around with the different views of the model to see if you can find other structural differences that may have resulted in the functional effects observed. Use this information to inform your own hypotheses as to how this mutation affects both the structure and function of the protein.

Conclusion

As can be seen in the models illustrated above, the most common documented mutations in the CYP2C9 gene do not appear to cause any major structural changes to the protein as a whole or to the domains the mutations fall within. There are some slight differences in the angles at which different domains sit from one another - which may be due to imperfections in the model or inconsistencies in viewing angle, though may also hold some biological relevance - but the overall shape is conserved. The Ile359Leu and Arg144Cys mutations (both illustrated above) are known to have clinical significance in the warfarin drug response, so there must be something going on at the molecular level that I am not able to see in these macrostructural models. It could be that the slight changes in angles is just enough to

alter binding ability, or perhaps the residues that were replaced are critical for ligand binding at the atomic level (Mesecar et al., 1997).

Although not planned, this example highlights one of the limitations of protein modeling such that the link between structure and function is not always obvious. Robetta provides data only to the level of amino acid positions, so further investigation, likely through experimental rather than predictive methods, could be needed to determine the exact source of the functional changes seen with these mutations. Some of the amino acid substitutions represented by the mutations modeled here contained significant changes in amino acid property such as the switch from basic, aliphatic arginine to pH-neutral, aromatic tryptophan in the Arg335Trp mutation; so this may be a case where the functional consequences are more due to the properties than the conformation of the mutated site being altered. It is also possible that the functional consequences of missense SNPs as a whole are not very well elucidated by these predictive models, though this can in no way be confirmed by the very small sample size of this project.

References

Johnson, J., Caudle, K., Gong, L., Whirl-Carrillo, M., Stein, C., Scott, S., Lee, M., Gage, B., Kimmel, S., Perera, M., Anderson, J., Pirmohamed, M., Klein, T., Limdi, N., Cavallari, L., & Wadelius, M. (2017). Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. *Clinical Pharmacology & Therapeutics*, *102*(3), 397–404. https://doi.org/10.1002/cpt.668

Mesecar, A. D., Stoddard, B. L., & Koshland, D. E. (1997). Orbital Steering in the Catalytic Power of Enzymes: Small Structural Changes with Large Catalytic Consequences. *Science*, *277*(5323), 202–206. https://doi.org/10.1126/science.277.5323.202

Meyer, U. A. (2000). Pharmacogenetics and adverse drug reactions. *The Lancet*, *356*(9242), 1667–1671. https://doi.org/10.1016/S0140-6736(00)03167-6

Roden, D. M., Wilke, R. A., Kroemer, H. K., & Stein, C. M. (2011). Pharmacogenomics. *Circulation*, *123*(15), 1661–1670. https://doi.org/10.1161/CIRCULATIONAHA.109.914820

Wiedermann, C. J., & Stockner, I. (2008). Warfarin-induced bleeding complications—Clinical presentation and therapeutic options. *Thrombosis Research*, *122*, S13–S18. https://doi.org/10.1016/S0049-3848(08)70004-5