Ronit Jain

CS 279

Homology Modeling and Validation of the Structure of Casbene Synthase

Introduction

Casbene synthase is an enzyme found in plants from the Euphorbicae family that catalyzes the cyclization of geranylgeranyl diphosphate (GGP) to casbene.[1] As part of a diverse class of 20-carbon molecules known as diterpenes, casbene has antifungal, antibacterial, and antiinflammatory properties, thereby serving as a crucial biomolecule for pharmaceutical and biomaterial development.[2] The overall reaction facilitated by casbene synthase is:

$$geranylgeranyl\ diphosphate\ \Leftrightarrow casbene + diphosphate$$

Despite its critical role in the synthesis of casbene from the precursor GGP, casbene synthase's structure has not yet been solved experimentally, representing a significant research gap in the field. As such, my project aims to utilize homology modelling to construct a model of the three-dimensional structure of casbene synthase and then cross-validate the structure with known biochemical data on the enzyme. In terms of the scope of the project, I plan to model one specific strain of casbene synthase isolated from the plant *Euphorbia pekinesis* (Peking spurge) and will focus on ligand docking experiments for the validation component of my project.

Background

Determining the structure of casbene synthase is an important problem to tackle because models of the enzyme can subsequently be used for bioengineering studies that aim to introduce specific mutations to the active site to increase the enzyme's efficiency. Under natural conditions, organic synthesis of casbene yields only small quantities of the molecule, thereby bottlenecking the manufacturing of medicinal compounds that rely on casbene as a raw material.[3] As such, designing a version of the enzyme that has enhanced catalytic activity will allow for increased availability of casbene, which has promising implications for the economic efficiency of biopharmaceutical development. Beyond its applications in industry, constructing a model of casbene synthase will elucidate the exact mechanisms underlying the biochemical pathway involved in diterpene synthesis, which is still poorly understood.[4]

Previous efforts on characterizing the structure of casbene synthase have primarily involved biochemical experiments as opposed to computational ones. In particular, site-directed mutagenesis has been leveraged to determine key residues necessary for the catalytic activity of the enzyme. For instance, one study engineered mutant versions of the enzyme with single amino acid changes and then observed the binding activity of the enzyme to the ligand, shedding light on which amino acids might be present in the active site.[5] Other work on casbene synthase has focused on identifying genes that code for the enzyme as well as methods of overexpressing those genes in yeast to improve biosynthesis of casbene.[6][7]

Methods

To begin the process of homology modelling, I looked up casbene synthase in the NCBI database and found the entry corresponding to the casbene synthase strain that had been isolated from *Euphorbia pekinesis.* Within the NCBI entry, I located the FASTA sequence of the protein, which I then inputted into BLAST to find other proteins in the Protein Data Bank (PDB) with highly similar sequences that could potentially serve as strong template candidates. The top hit outputted by BLAST was henbane vestipiradiene synthase (HVS) with a query cover of 88% and an E-value of 9e-155. To verify the quality of the experimentally-solved crystal structure for this enzyme, I used ProCheck to examine its Ramachandran Plot for any stereochemical outliers. Next, I inputted the PDB file of HVS into SWISS-MODEL as the template and the FASTA sequence of casbene synthase as the target sequence. Using the template I had identified, SWISS-MODEL subsequently generated a three-dimensional structure of casbene synthase, which I downloaded as a PDB file for further analysis.

To validate this homology model, I utilized SWISS-DOCK to conduct ligand docking of the homology structure to its substrate, GGP, whose structure I downloaded from the ZINC ligand database. All hydrogens were added to the ligand as well as the homology structure. The receptor-ligand complex for each binding pose was then visualized using UCSF Chimera. Once docking of the homology model had been completed, I wanted to validate the model against published biochemical data on the enzyme. One particular study used site-directed mutagenesis to generate mutant strains of casbene synthase (specifically, D355E, D356E, D359E) which targeted a conserved aspartate-rich region of the enzyme (DDTID, residues 355-359) that is found across other known synthases.[5] The experimental results revealed that D355E and D356E had significantly lower binding to the ligand while D359E had no such change in binding, suggesting that the aspartate residues at positions 355 and 356 are necessary for the enzyme's catalytic activity.[4] As such, I decided to computationally recreate those same mutants and then test whether their docking to the ligand differed from the wild-type enzyme. To do so, I conducted site-directed mutagenesis on the homology model using PyMOL, ultimately generating three protein structures which had the aspartate residues at positions 355, 356, and 359 changed to glutamate. Then, I repeated the same procedure outlined above to conduct ligand docking of the mutated enzymes with GGP.

For each binding pose, SWISS-DOCK outputs a ΔG value which indicates the energetic favorability of the binding—however, I converted these ΔG values into K_D values, which is the ligand concentration at which half the ligand binding sites of the enzyme are occupied and therefore represents a more useful value for the purposes of this project. To do so, I used the following equation relating K_D to ΔG: $K_D = e^{\frac{\Delta G}{RT}}$, where $R = 8.314\, J\, mol^{-1}\, K^{-1}$ and $T = 298.15\, K$.

Results

Homology Model Construction

The homology model generated via SWISS-MODEL using HVS as a template was visualized in PyMOL and is reproduced below. The aspartate-rich conserved region is colored red for reference.

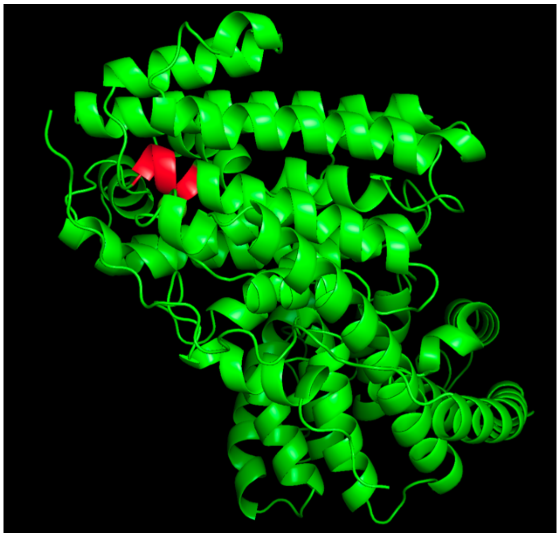

Figure 1: Cartoon View

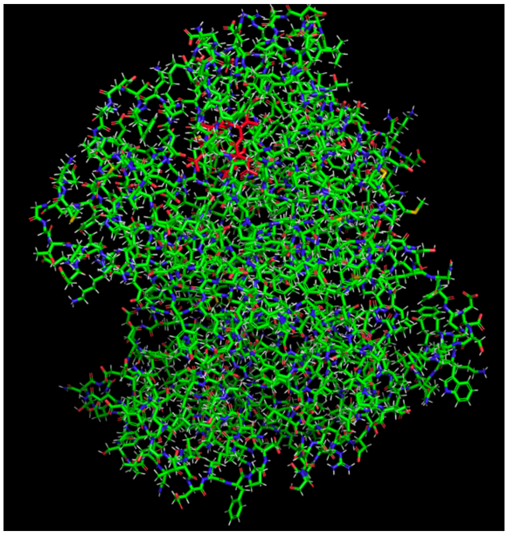

Figure 2: Licorice View

Wild-Type Enzyme Ligand Docking

Below is data from SWISS-DOCK regarding the top eight binding poses for the docking of the homology model with the ligand GGP, along with their associated ΔG values which I have also converted into K_D values:

Pose	ΔG (kcal/mol)	K_D (M)
1	-12.91	3.44×10^{-10}
2	-12.83	3.94×10^{-10}
3	-12.33	9.15×10^{-10}
4	-12.05	1.47×10^{-9}
5	-11.39	4.47×10^{-9}
6	-11.39	4.47×10^{-9}
7	-11.38	4.55×10^{-9}
8	-11.38	4.55×10^{-9}

The ligand-receptor complex for the top three binding poses were visualized using UCSF Chimera, with the aspartate-rich conserved region colored red.

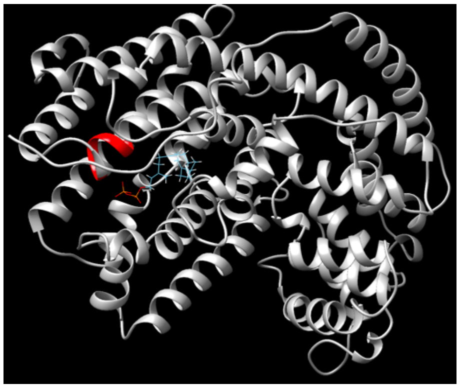

Figure 3: Pose 1

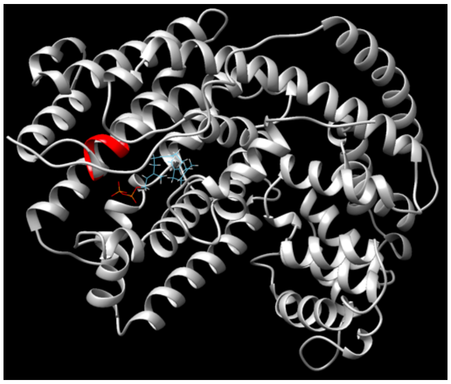

Figure 4: Pose 2

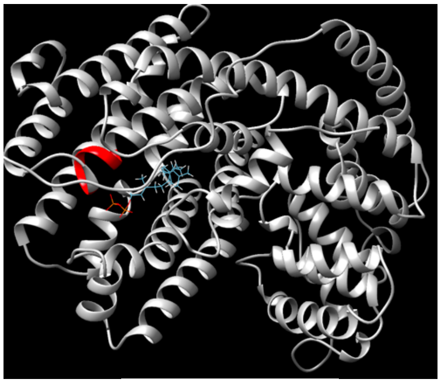

Figure 5: Pose 3

D355E Mutant Enzyme Ligand Docking

Below is data from SWISS-DOCK regarding the top eight binding poses for the docking of the D355E mutant enzyme with the ligand GGP, along with their associated ΔG values which I have also converted into K_D values:

Pose	ΔG (kcal/mol)	K_D (M)
1	-11.29	5.30×10^{-9}
2	-11.29	5.30×10^{-9}
3	-11.25	5.67×10^{-9}
4	-11.25	5.67×10^{-9}
5	-11.23	5.86×10^{-9}
6	-11.23	5.86×10^{-9}
7	-11.20	6.16×10^{-9}
8	-11.20	6.16×10^{-9}

The ligand-receptor complex for the top three binding poses were visualized using UCSF Chimera, with the aspartate-rich conserved region (now EDTID in the mutant enzyme) colored red.

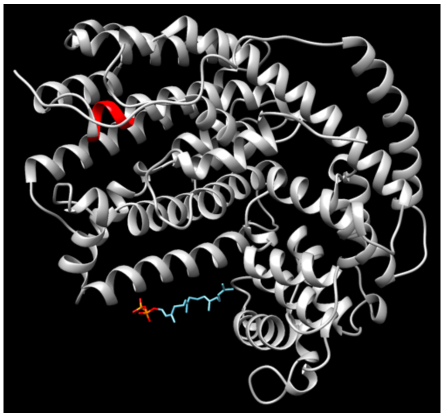

Figure 6: Pose 1

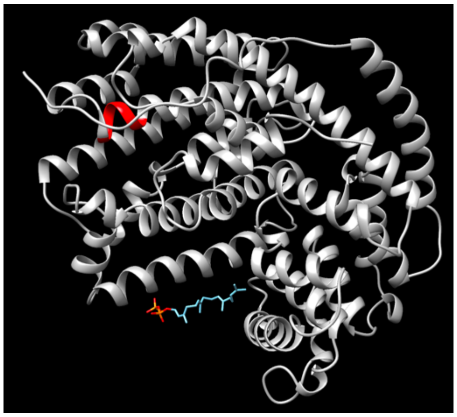

Figure 7: Pose 2

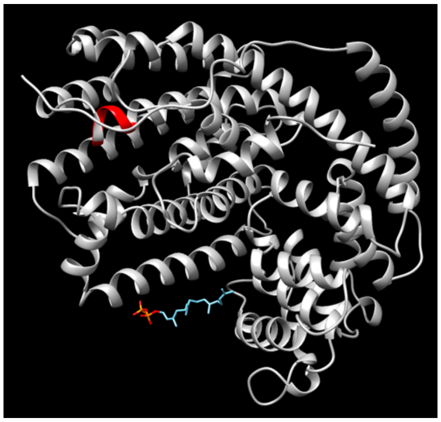

Figure 8: Pose 3

D356E Mutant Enzyme Ligand Docking

Below is data from SWISS-DOCK regarding the top eight binding poses for the docking of the D356E mutant enzyme with the ligand GGP, along with their associated ΔG values which I have also converted into K_D values:

Pose	ΔG (kcal/mol)	K_D (M)
1	-11.07	7.68×10^{-9}
2	-11.07	7.68×10^{-9}
3	-11.03	8.21×10^{-9}
4	-11.03	8.21×10^{-9}
5	-11.03	8.21×10^{-9}
6	-11.03	8.21×10^{-9}
7	-10.97	9.09×10^{-9}
8	-10.89	1.04×10^{-8}

The ligand-receptor complex for the top three binding poses were visualized using UCSF Chimera, with the aspartate-rich conserved region (now DETID in the mutant enzyme) colored red.

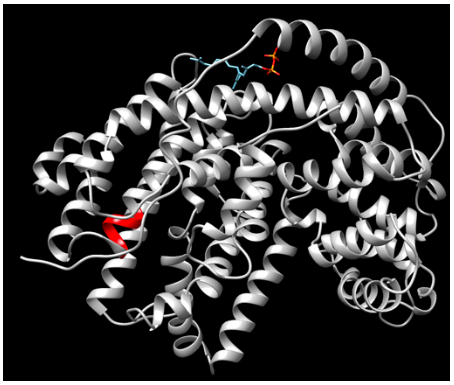

Figure 9: Pose 1

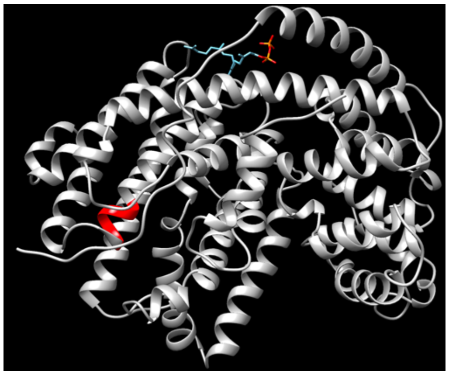

Figure 10: Pose 2

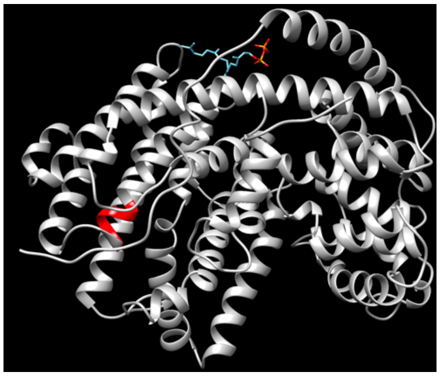

Figure 11: Pose 3

D359E Mutant Enzyme Ligand Docking

Below is data from SWISS-DOCK regarding the top eight binding poses for the docking of the D359E mutant enzyme with the ligand GGP, along with their associated ΔG values which I have also converted into K_D values:

Pose	ΔG (kcal/mol)	K_D (M)
1	-12.43	7.73×10^{-10}
2	-12.35	8.85×10^{-10}
3	-12.35	8.85×10^{-10}
4	-11.20	6.16×10^{-9}
5	-10.96	9.24×10^{-9}
6	-10.92	9.89×10^{-9}
7	-10.92	9.89×10^{-9}
8	-10.87	1.08×10^{-8}

The ligand-receptor complex for the top three binding poses were visualized using UCSF Chimera, with the aspartate-rich conserved region (now DDTIE in the mutant enzyme) colored red.

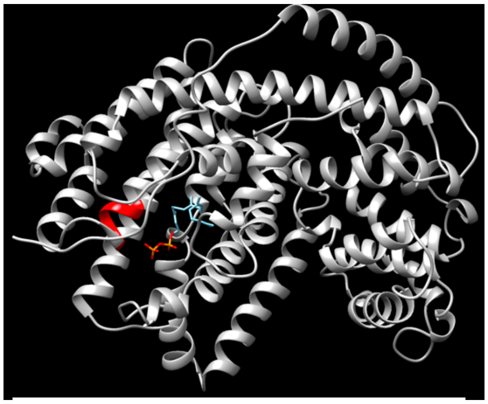

Figure 12: Pose 1

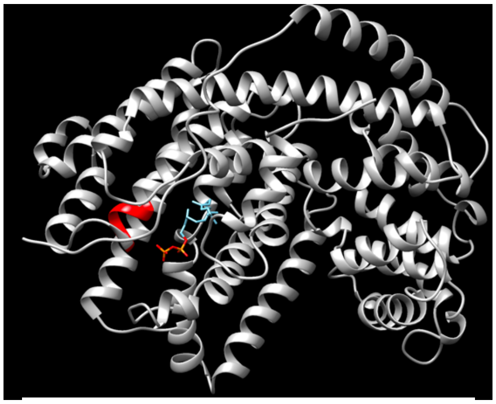

Figure 13: Pose 2

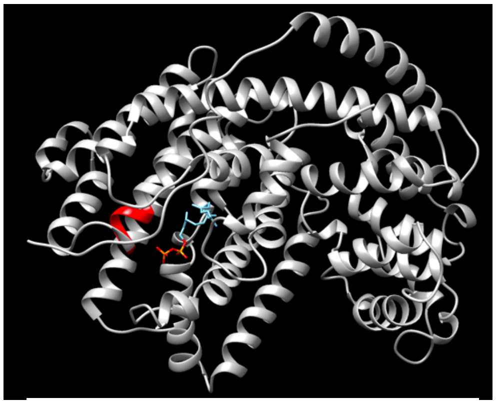

Figure 14: Pose 3

Discussion

Overall, I found that the ligand docking results aligned quite well with the biochemical data on the enzyme. Firstly, I noted that the top-scoring pose for the ligand docking of the wild-type enzyme (homology model) showed the ligand binding to a region immediately adjacent to the aspartate-rich region. Since the experimental study concluded that two of the aspartates in the aspartate-rich region (D355 and D356) are critical for the enzyme's catalytic activity, it makes sense that the aspartate-rich region would be included in the active site to which the ligand binds. Moreover, the second and third-highest scoring poses also showed the ligand binding to the same pocket of the enzyme that included the aspartate-rich region, thereby bolstering my confidence in the validity of the homology model.

The D355E mutant enzyme binding poses were very distinct from the wild-type enzyme binding poses. As can be seen in the UCSF Chimera visualizations, the ligand is no longer predicted to bind to the aspartate-rich region but rather to the opposite end of the enzyme. Interestingly, the ligand does not seem to be embedded within the enzyme in any sort of binding pocket—instead, it appears to be superficially attached to the surface of the enzyme. This does not seem physically realistic, since most active sites are buried within the enzyme as opposed to on the surface of the enzyme, so I suspect that this binding pose might not actually occur naturally. It is also worth noting that the K_D value for this binding pose (5.30×10^{-9}) is quite a bit higher than the K_D value for the binding pose of the wild-type enzyme to the ligand (3.44×10^{-10}), which indicates that the binding affinity of this particular pose is not as high as that of the wild-type. It seems, then, that introducing the mutation might reduce the ability of the enzyme to bind to its substrate. This corresponds with the biochemical data which revealed a significant increase in K_m value from 3.1 μM to 32 μM following the introduction of the mutation, indicating that higher substrate concentrations are now needed to achieve maximum reaction velocity. Overall, I think the fact that the top three binding poses did not involve the aspartate-rich region suggests that the ligand is no longer able to bind to that pocket when the enzyme contains the D355E mutation, corroborating the experimental data.

Similar to the D355E mutant, the D356E mutant also had binding poses that were markedly different from the wild-type enzyme. Once again, the ligand was no longer predicted to bind to the aspartate-rich region but rather to a completely different region of the enzyme, potentially indicating that the mutation prevents the ligand from binding to the aspartate-rich pocket. Notably, the K_D value for the top-scoring pose for the D356E mutant (7.68×10^{-9}) was much higher than the K_D value for the top-scoring pose for the wild-type enzyme (3.44×10^{-10}), which could suggest that the D356E mutation makes the enzyme unable to bind to the substrate with the same affinity as in its wild-type form. This fits well with experimental data which found that the D356E mutant had lowered catalytic activity (K_m of 12.5 μM).

The D359E mutant binding poses were all localized to the aspartate-rich conserved region and strongly resembled the binding poses from the wild-type enzyme ligand docking, which leads me to believe that this mutation does not significantly impact the enzyme's ability to bind to its substrate. The K_D value for the top-scoring pose for D359E (7.73×10^{-10}) was also comparable to the K_D value for the top-scoring pose for the wild-type enzyme (3.44×10^{-10}), which further suggests that the binding affinity of the ligand to the enzyme is not affected. The experimental biochemical data supports this conclusion, as the K_m value of the D359E mutant enzyme was virtually unchanged compared to the wild-type enzyme (3.2 μM vs 3.1μM). I was curious as to why this mutation did not affect the ligand docking, so I examined the aspartate residue within the homology model in PyMOL and found that its side chain was oriented away from the ligand binding pocket, whereas the side chains of the other two aspartates (D355 and D356) were positioned within the binding pocket. This difference in orientation is because the all three aspartates are part of an alpha helix. As such, I suspect that the D359 aspartate side chain does not actually interact with the substrate due to its physical orientation, whereas the other two aspartates (D355 and D356) are critical for binding because their side chains protrude into the binding cavity.

Additionally, I was surprised to see that many of the docking results outputted binding poses with very similar binding affinities. Oftentimes, SWISS-dock grouped these binding poses together in a "cluster," but the existence of so many different binding poses with similar affinities suggests that the ligand might be spending time in multiple binding conformations. However, it is also important to keep in mind that SWISS-DOCK uses an empirical scoring function that only roughly approximates binding free energy based on what scientists believe makes a ligand-receptor interaction energetically favorable (i.e. hydrogen bonds, contact between hydrophobic atoms, etc.). As such, the similarity in ΔG values between different poses might just be a result of this approximation and therefore cannot be taken as completely authoritative.

Moreover, upon examining the 3D structure of HVS (the template), I found that the regions where the amino acid sequences of the template and the target matched completely had the exact same secondary structure while regions where the template did not have any corresponding sequence often had quite different structures. Upon further reflection, I think this makes sense given how SWISS-MODEL conducts protein modelling—regions in the target with the exact same sequence as the template simply have their structure imported from the template, whereas regions with differences are constructed using loop modelling which searches for similar fragments in the PDB database. However, this also made me wonder whether another choice of template might have subtly changed my results. Another approach that might work could involve using molecular dynamics simulation to visualize the process by which the protein folds. This simulation process would not be dependent on a template but rather on a molecular mechanics force field which estimates the potential energy associated with each arrangement of atoms in the protein. While this could circumvent the issue of template and also provide insight into the process of folding, there is the downside of needing increased computational power to do the simulation. Moreover, the molecular dynamics simulation might not be able to predict the formation of disulfide bonds since covalent bonds cannot be altered during the simulation. Another possible approach might be to use Phyre2 in its intensive mode, which would allow for the construction of a model using multiple templates that match different parts of the sequence. Due to time constraints, I was unable to play around with Phyre2, but this could lead to a more accurate model since we are no longer relying on just one template to guide the construction process. I'd be curious to see how similar the Phyre2 model is to the SWISS-MODEL one.

From these experiments, I learned that homology modelling in tandem with known experimental data can be a powerful way to validate a potential 3D-structure of a protein of interest. In particular, I found that the mutants I generated via site-directed mutagenesis had very similar results to the biochemical data in the ligand docking experiments. I also realized that choosing the best binding pose can at times be a subjective exercise—although binding affinities are given for each pose, one must also visually examine the poses to see if they seem plausible. Finally, I learned that the choice of a template for homology modelling is critical and must have sufficient similarity to the target sequence in order for the modelling to be successful—I had previously chosen a protein that only had sequences with ~20% sequence similarity, which resulted in quite inaccurate models. Overall, this project fits into the broader theme of this class because it explores the ways that computational and experimental approaches to protein modelling can complement one another. Moreover, my project explores the impacts of making strategic modifications to a protein on its ability to bind to its substrate, which is a key topic in computational biology with direct applications in designing more efficient enzymes for use in agriculture, biotechnology, and pharmaceutical development.

Future Work

In the future, I'd like to investigate how we might engineer a mutated version of the casbene synthase that binds to its substrate more tightly, potentially leading to an increase in its catalytic activity. I would be interested in introducing single amino-acid changes to the putative active site, identified in this project via ligand docking, and then observe their effect on the binding affinity of the ligand. For instance, since GGP is an amphipathic molecule, it might be worthwhile to try introducing more polar residues to the region where the polar head of the molecule binds and more nonpolar residues to the region where the nonpolar body attaches to see if this affects the energetic favorability of the binding. These computational results can then be verified experimentally.

I would also like to perform homology modelling for other strains of casbene synthase isolated from various plant species. Currently, NCBI has FASTA sequences for over ten strains of casbene synthase, so it would be interesting to see which structural features are conserved across these strains. I'd also be interested to see if certain strains seem to have higher binding affinities for GGP than others, which might provide insight into what particular secondary structure features facilitate greater catalytic activity.

Finally, I am curious to see whether docking with flexible residues might change the results of the ligand docking experiments. For this project, I did not set any of the residues in the receptor as flexible, since SWISS-DOCK does not offer the functionality of setting specific side chains as flexible. However, it could be worthwhile to try the same docking experiments in AutoDock Vina, which does have this functionality, and observe the changes in binding pose/energetic favorability that occur as a result of flexible side chains. Furthermore, I could also try to change which particular side chains are flexible to ascertain whether that has any effect on the binding.

References

[1] Mau C. et al. "Cloning of casbene synthase cDNA: evidence for conserved structural features among terpenoid cyclases in plants." *Proceedings of the National Academy of Sciences*, vol. 91. 25 April, 1994.

[2] De Almeida, N.P. et al. "Monitoring casbene synthase in *Jatropha curcas* tissues using targeted proteomics." *Plant Methods*, vol. 17. 6 February, 2021.

[3] Luo, D. et al. "Oxidization and cyclization of casbene in the biosynthesis of *Euphorbia* factors from mature seeds of *Euphorbia lathyris.*" *Proceedings of the National Academy of Sciences*, vol. 113. 9 August, 2016.

[4] Callari, R. et al. "Dynamic control of *ERG20* and *ERG9* expression for improved casbene production in *Saccharomyces cerevisiae.*" *Frontiers in Physiology.* 1 November, 2018.

[5] Huang, K. et al. "Overexpression, single-step purification, and site-directed mutagenetic analysis of casbene synthase." *Archives of Biochemistry and Biophysics,* vol. 352. 1 April, 1998.

[6] Hill, A.M. et al. “High level expression of Ricinus communis casbene synthase in Escherichia coli and characterization of the recombinant enzyme.” *Arch Biochem Biophys*. 15 Dec, 1996.

[7] Ding, B. et al. “Characterization of the geranylgeranyl diphosphate synthase gene in *Acyrthosiphon pisum* and its association with carotenoid biosynthesis.” *Frontiers in Physiology.* 12 November, 2019.