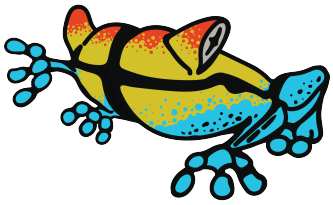
Synthetic Biobots

Docking Methods



SYNTHETIC
BIOBOTS

Author: Gerardo Cendejas-Mendoza

iGEM Design League

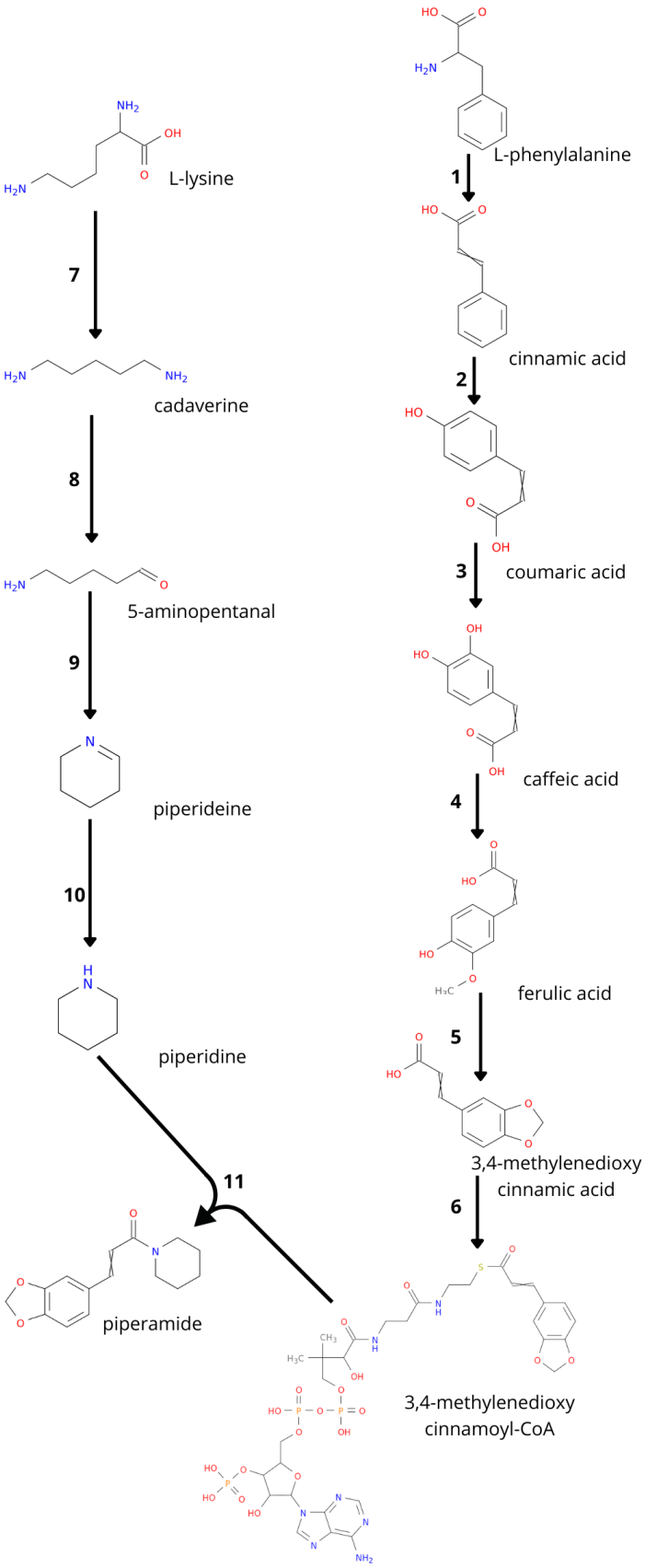
Season 2022

Introduction

In this document is shown the docking methodology followed by the team Synthetic Biobots during the iGEM Design League 2022 season. Molecular docking is a process to couple proteins with their ligands; for this project molecular docking was performed in order to support the decision of using the proteins that were selected as part of the biosynthetic pathway of piperamide. The different steps necessary for its production are shown in the figure shown in the next page; in this figure all the steps required are marked with a number, this number is an identifier for the enzyme required to perform that step.

From all the enzymes used in the pathway, only 8 require a docking step for different reasons, these reasons can be that it is an enzyme that has not been isolated and even though a previous annotation step indicates it performs that reaction, molecular docking can provide better support for their selection and activity. Another reason to perform molecular docking is when the protein is known and has been isolated but we are using a molecule that is not the original ligand of the protein, in this case molecular docking can help to elucidate if the protein is capable of performing the reaction with a molecule with similar structure to its original ligand. The enzymes that are subject of molecular docking and the genes that code for these enzymes and will be studied are: 1 (Pn8.2617), 2 (Pn2.84), 4 (Pn1.1317), 5 (Pn3.4770, Pn7.1626), 6 (Pn16.1237, Pn16.1198), 8 (Pn4.3222), 9 (Pn2.2377) and 10 (2cwh).

For each of these genes it was performed a differential exon usage analysis between four different *Piper nigrum* tissues: fruit at 20 days stage, fruit at 40 days stage, leaf and panicle. This analysis can help us to visualize the expression level for each exon of the genes, this information allows us to select the exons of the gene to include in the modeling and in the following steps. This analysis was performed by using the Bioconductor library "DEXSeq", to access the code used for this methodology please visit the GitHub repository of the team for iGEM Design League 2022 season, the Exon_usage directory.[10;3;27] The STAR ultrafast universal RNA-seq aligner on the Galaxy project platform mapped the RNA-seq reads to the reference genome of *Piper nigrum* prior to the differential exon usage analysis.[12;7;6] Only results of relevance for the modeling and docking methods are shown in this document.

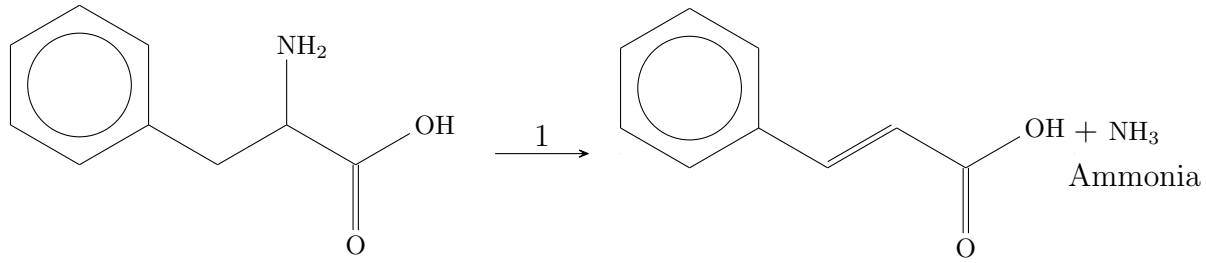


Biosynthetic pathway of piperamide.

1 Docking

1.1 Pn8.2617

The first reaction in the biosynthetic pathway of piperamide is performed by the enzyme phenylalanine ammonia-lyase, which is part of the phenylpropanoid biosynthesis pathway and catalyzes the reaction shown in Eq. 1. The gene Pn8.2617 was selected as the best candidate for enzyme 1, this selection was based on the fact that it obtained the highest score (1170.9) over the threshold (524.93) for KEGG Orthology ID K10775, with E-value $\approxeq 0$. This annotation was made with KofamKOALA.[4] With the exon usage analysis no relevant results were found.



Phenylalanine Cinnamic acid (1)

The gene contains two exons that were considered when modeling the three-dimensional structure of the protein. The structure modeling was made by using an homology based methodology "SWISS-MODEL".[36;5] This methodology is template based, in this case the template selected was a phenylalanine ammonia-lyase from *Petroselinum crispum* with PDB ID: 1w27; this protein has a 79.97% sequence identity to our query protein Pn8.2617 which makes it an appropiate template. The template was found with HHblits, a protein sequence searching algorithm that uses hmm-hmm alignment.[26] The model (Fig. 1.1.1) is found in a homo-tetrameric state, as the template is; it has a GMQE value of 0.84 and a QMEANDisCo Global value of 0.81 $\pm$ 0.05 (Fig. 1.1.2). The QMEANDisCo value is a composite score value for single model quality estimation, this value is a function of underlying single model scores employing statistical potentials of mean force and a distance constraint based on the known structure of homologous proteins. QMEANDisCo value is

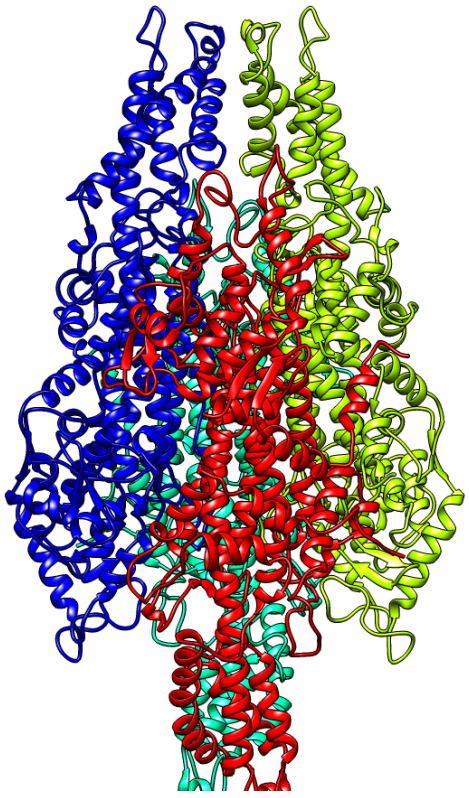


Figure 1.1.1: Pn8.2617 three-dimensional model.

calculated for every residue of the protein and the QMEANDisCo Global value is the average of these per residue scores, which aims to be a good approximation of the global overall quality of the model.[29] The ligand 3D conformer structure was obtained from PubChem with CID: 6140.

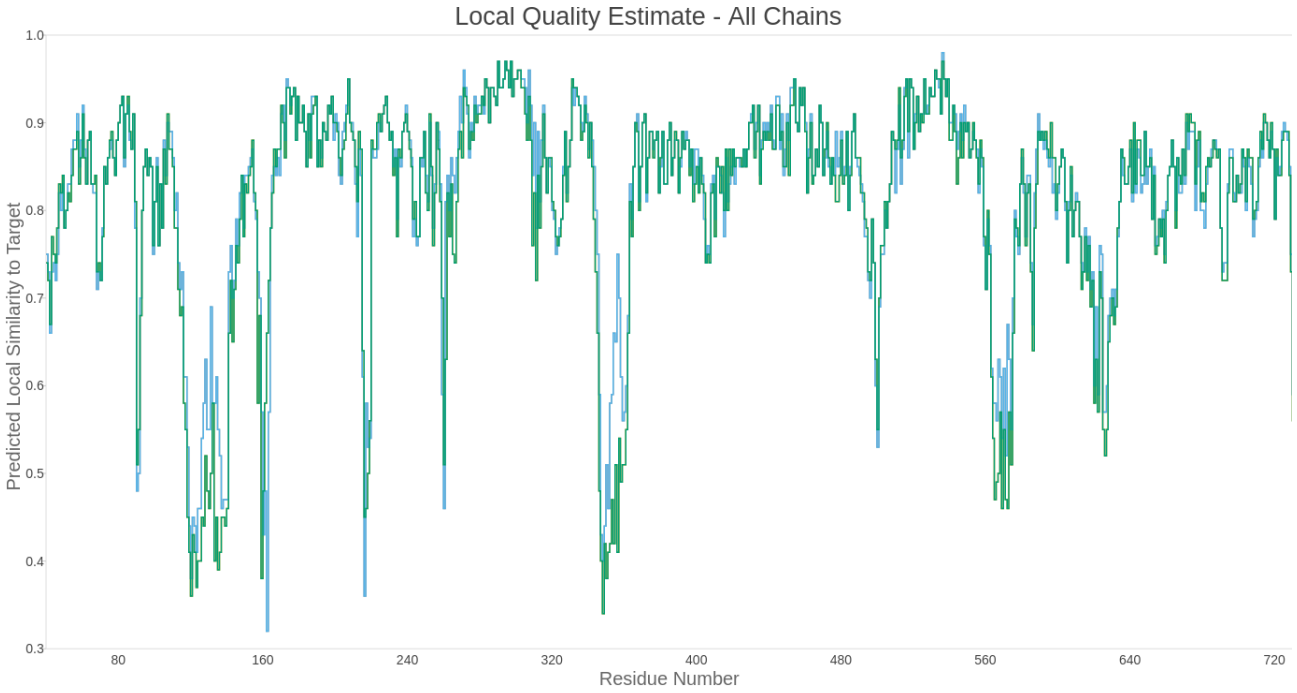


Figure 1.1.2: Pn8.2617 model QMEANDisCo.

Before docking was performed the protein and the ligand were energy minimized. The protein structure was minimized using UCSF ChimeraX software[24;11] and the steepest descent algorithm for 1000 steps with a step size of 0.02 Å; the force field AMBER ff14SB was used for standar residues, for non standard residues the semi-empirical AM1-BCC model that uses an additive bond charge corrections was selected since it has shown higher success rate than other force fields in docking methods.[13;14;34] In the same way the ligand was energy minimized by using using the python library "rdkit" and the MMFF94 force field implemented within it.[16;32]

The preparation of the molecules for docking was made on AutoDockTools and Autodock Vina version 1.2.3 docked the protein and the ligand.[23;8;33] The grid box for the docking was selected based on the conserved domains found with the NCBI conserved domain search interface; we found the tetramer interface and active site conserved domains for phenylalanine ammonia-lyase from the CDD database.[17;18]. CASTp3 found the protein's pockets, which are cavities on its surface into which solvent can enter; CASTp3 uses computational geometry to find these pockets mainly using Delaunay triangulation, alpha shape, and discrete flow.[31] As a result of molecular docking with Vina, nine different conformations of the ligand in the protein were found, these results are shown in Table 1.

The docked positions of the phenylalanine in the Pn8.2617 structure interacted with different aminoacid residues of the protein (Fig. 1.1.3). Some of these aminoacids were found in the conserved domain for the active site of the enzyme, like Phe 129, Leu 147, Leu 270, Asn 274, Tyr 365, Arg 368 and Asn 398; the rest of the aminoacids were not marked as conserved in the active site domain but were present in the pockets found by CASTp3: Phe 150, Leu 221, Phe 414, Lys 470, Glu 498, Asn 501 and Gln 502.

Table 1: Docking results of phenylalanine in Pn8.2617

Mode	Affinity (kcal/mol)	Dist. from best mode RMSD l.b	RMSD u.b
1	-6.278	0.0	0.0
2	-6.126	2.278	2.795
3	-6.098	1.934	2.262
4	-6.095	1.773	1.908
5	-6.013	1.735	2.332
6	-5.850	2.656	3.643
7	-5.847	1.749	2.333
8	-5.700	2.823	3.134
9	-5.670	3.526	4.412

A further analysis was carried out to elucidate the protein-ligand interactions of the best conformation obtained, which has a binding energy of -6.278 kcal/mol. A search for hydrogen bonds was performed using UCSF ChimeraX allowing relax constraints of 0.4 Å and 20 degrees, the bonds are shown as pseudobonds since they represent significant interactions between atoms other than covalent bonds, also because of the relax constraint these predictions can represent bonds within the tolerance values but not meeting the precise criteria.[24;11;21] Four pseudobonds were found between the ligand and two aminoacids of the protein, one pseudobond with Asn 501 and a distance of 2.022 Å and three with Arg 368 with distances of 2.240 Å, 2.322 Å and 2.438 Å. In Fig. 1.1.4 this pseudobonds and the ligand in its pocket are shown, the surface of the protein is colored based on its hydrophobicity using Kyte-Doolittle scale from lower (blue) to higher hydrophobicity (red).

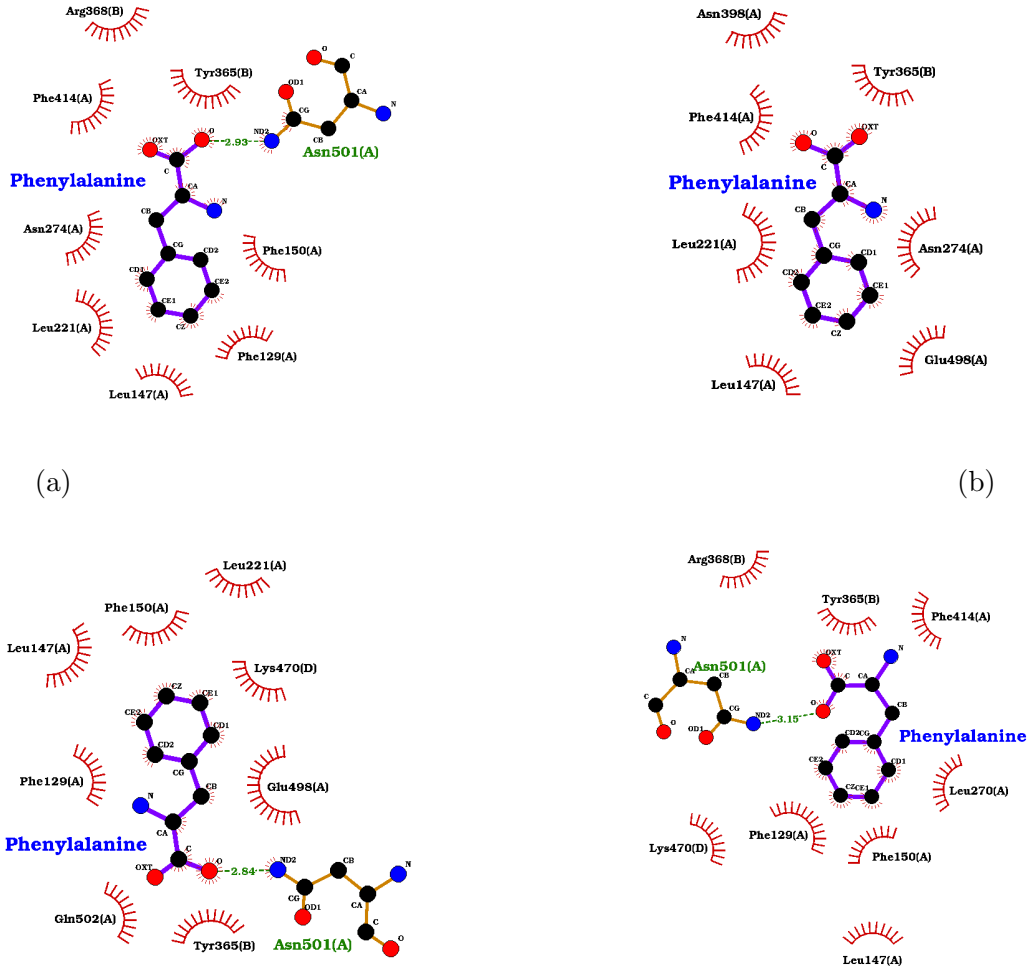


Figure 1.1.3: Interactions of phenylalanine with Pn8.2617. Best four poses are shown in order a-d.

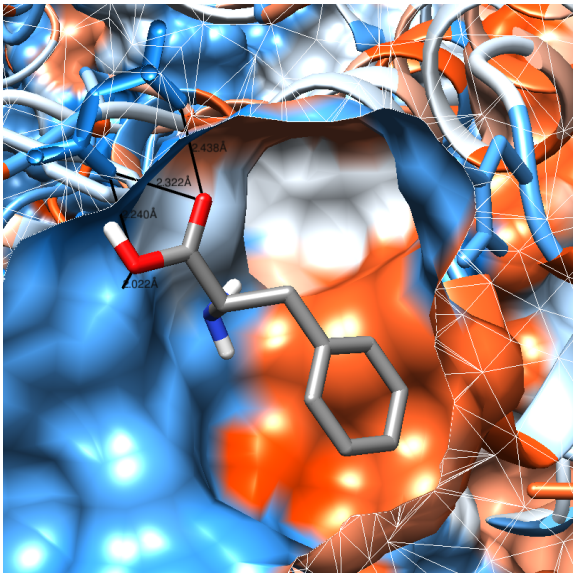
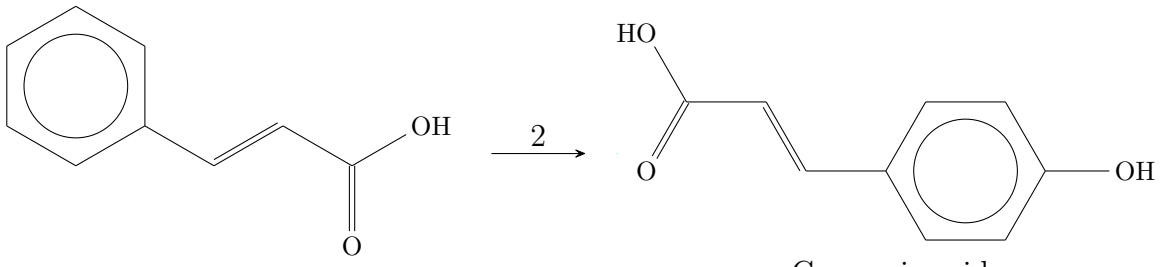


Figure 1.1.4: Phenylalanine pseudobonds with Pn8.2617.

1.2 Pn2.84

The second reaction of the pathway (Eq. 2) is catalyzed by a protein from the Cytochrome P450 Monooxigenase family called trans-cinnamate 4-hydroxylase, this enzyme is also part of the phenylpropanoid biosynthesis pathway. The gene Pn2.84 was selected as the best candidate for enzyme 2 based on the fact that it obtained the highest score (985.2) over the threshold (555.2) for KEGG Orthology ID K00487, with E-value = 2.5×10^{-297}. This annotation was made with KofamKOALA.[4] With the exon usage analysis no relevant results were found.



Cinnamic acid Coumaric acid (2)

The gene contains four exons that were considered when modeling the three-dimensional structure of the protein. The structure modeling was made by using an homology based methodology "SWISS-MODEL".[36] The selected template was a cinnamate 4-hydroxylase from *Sorghum bicolor* with PDB ID: 6VBY; this protein has a 77.91% sequence identity to our query protein Pn2.84 which makes it an appropiate template. The template was found with HHblits, a protein sequence searching algorithm that uses hmm-hmm alignment.[26] The model (Fig. 1.2.1) is found in a monomeric state, as the template is; it has a GMQE value of 0.91 and a QMEANDisCo Global value of 0.91 ± 0.05 (Fig. 1.2.2).[29] The enzyme requires an heme molecule as a coenzyme for its correct catalytic activity; in the template a protoporphyrin IX containing Fe was found, which was used as a coenzyme for Pn2.84. The ligand 3D conformer structure was obtained from PubChem with CID: 5957728.

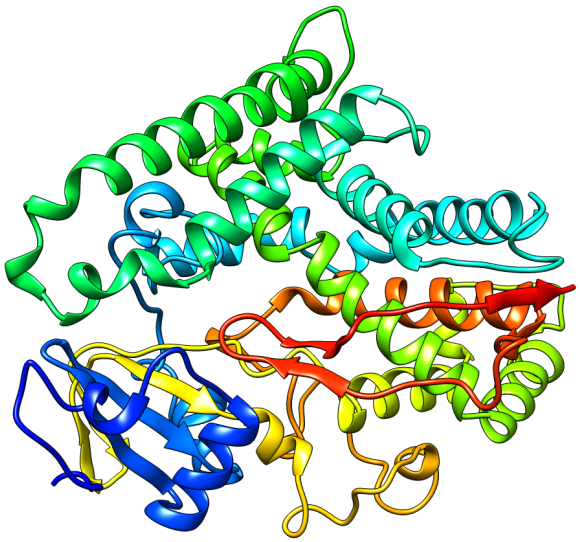


Figure 1.2.1: Pn2.84 three-dimensional model.

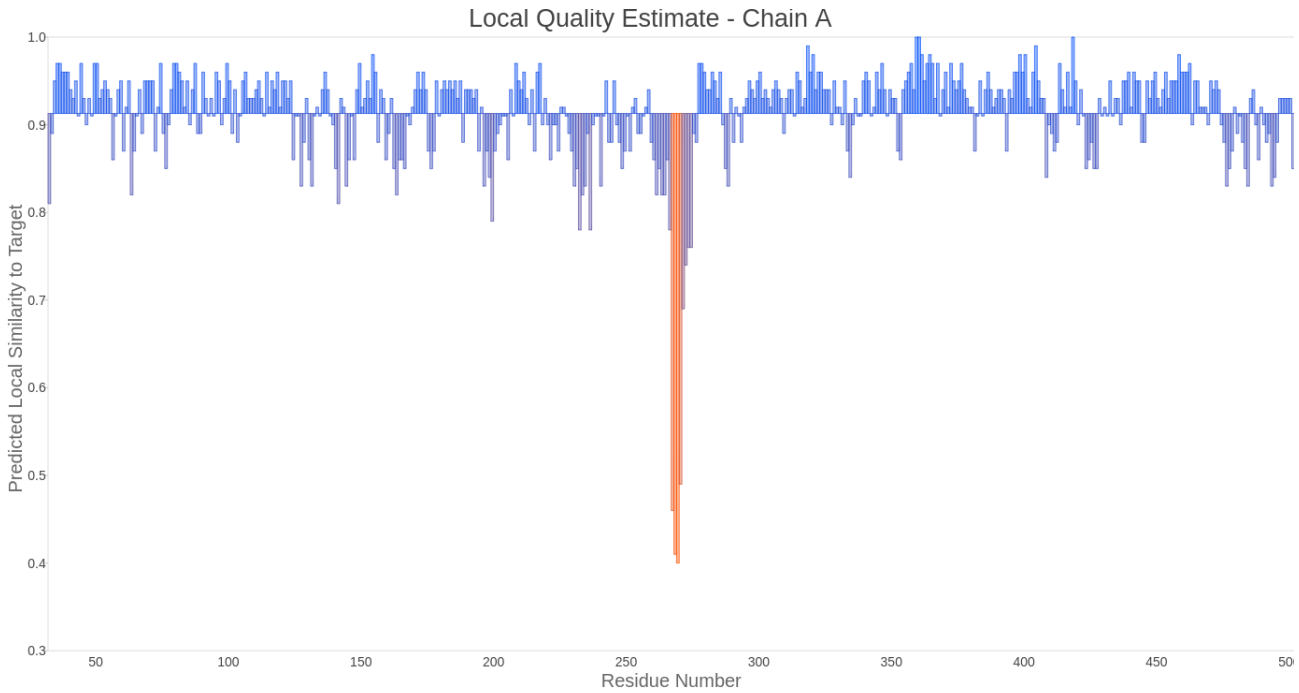


Figure 1.2.2: Pn2.84 model QMEANDisCo.

Before docking was performed the protein and the ligand were energy minimized. The protein structure was minimized using UCSF ChimeraX software[24;11] and the steepest descent algorithm for 1000 steps with a step size of 0.02 Å; the force field AMBER ff14SB was used for standar residues, for non standard residues the semi-empirical AM1-BCC model.[13;14;34] In the same way the ligand was energy minimized by using using the python library "rdkit" and the MMFF94 force field implemented within it.[16;32]

The preparation of the molecules for docking was made on AutoDockTools and Autodock Vina version 1.2.3 docked the protein and the ligand.[23;8;33] The grid box for the docking was selected based on the conserved domains found with the NCBI conserved domain search interface; we found the heme binding site and putative chemical substrate binding pocket conserved domains for trans-cinnamate 4-hydroxylase from the NCBI-Curated Domains database; additionally, active site residues were selected from the template structure, which included mainly Arg 213, Ser 214 and Gln 218.[17;18]. CASTp3 found the protein's pockets.[31] To perform an adequate docking of both the heme group and the cinnamate molecule a sequential docking workflow was used as described in Vass et al. 2012; in this methodology the protein and ligand were prepared (energy minimized) before the docking was performed, then the first molecule was docked to the protein, the best pose was merged with the protein and this complex was used as a the receptor to dock the next compound.

1.2.1 Protoporphyrin IX with Fe

The first molecule docked was the protoporphyrin IX with Fe, as a result of molecular docking with Vina, nine different conformations of the molecule in the protein were found, these results are shown in Table 2. The docked positions of the protoporphyrin IX with Fe in the Pn2.84 structure interacted with different aminoacid residues of the protein (Fig. 1.2.3). Some of these aminoacids were found in the conserved domain for the heme binding site of the enzyme, like Arg 101, Trp 126, Arg 130, Phe 138, Met 186, Ile 303, Ala 306, Ala 307, Thr 310, Thr 311, Leu 374, Val 375, His 377, Pro 439, Phe 440, Gly 441, Arg 445, Ser 446, Cys 447, Pro 448, Gly 449, Leu 452, and Ala 453; the rest of the aminoacids were not marked as conserved in the heme binding site domain but were present in the pockets found by CASTp3: Leu 91, Met 117, Val 118, Met 133, Ser 314, Ile 365, Ile 371 and Leu 457.

Table 2: Docking results of protoporphyrin IX (Fe) with Pn2.86

Mode	Affinity (kcal/mol)	Dist. from best mode RMSD l.b	RMSD u.b
1	-9.795	0	0.0
2	-8.72	0.7843	6.157
3	-8.471	2.831	7.051
4	-8.274	3.895	8.459
5	-7.935	2.005	4.936
6	-7.545	2.033	6.57
7	-6.66	3.496	8.093
8	-6.646	3.464	7.518
9	-6.506	2.739	6.189

A further analysis was carried out to elucidate the protein-ligand interactions of the best conformation obtained, which has a binding energy of -9.795 kcal/mol. A search for hydrogen bonds was performed using UCSF ChimeraX allowing relax constraints of 0.4 Å and 20 degrees.[24;11;21] Eight pseudobonds were found between the ligand and four aminoacids of the protein: three pseudobonds with Arg 101 with distances 2.033 Å, 2.169 Å and 2.206 Å; one pesudobond with Trp 126 of 2.197 Å; one with His 377 of 2.104 Å; and three with Arg 445 with distances 1.964 Å, 2.128 Å and 2.238 Å. In Fig. 1.2.4 this pseudobonds and the ligand in its pocket are shown, the surface of the protein is colored based on its hydrophobicity using Kyte-Doolittle scale from lower (blue) to higher hydrophobicity (red).

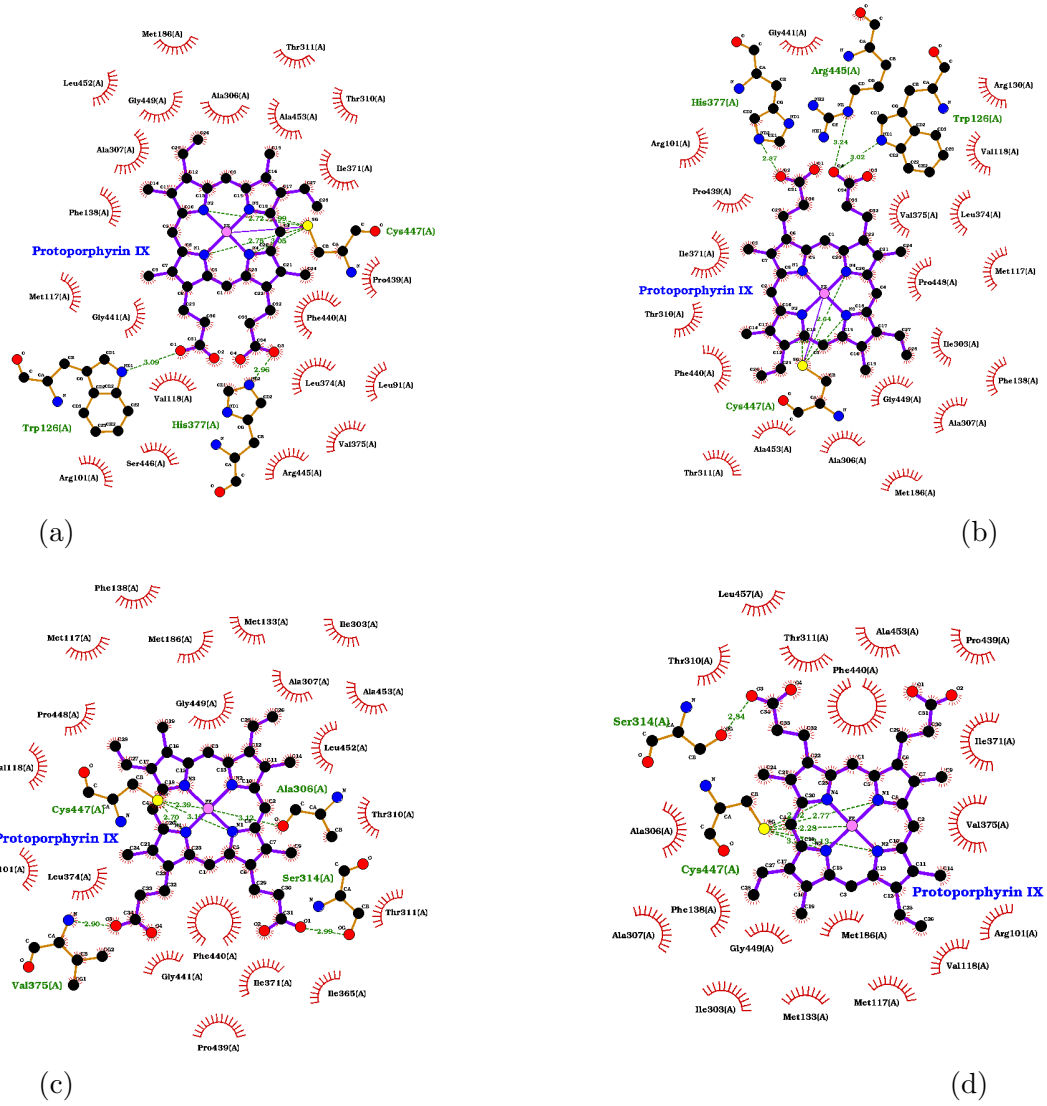


Figure 1.2.3: Interactions of protoporphyrin IX with Fe and Pn2.84. Best four poses are shown in order a-d.

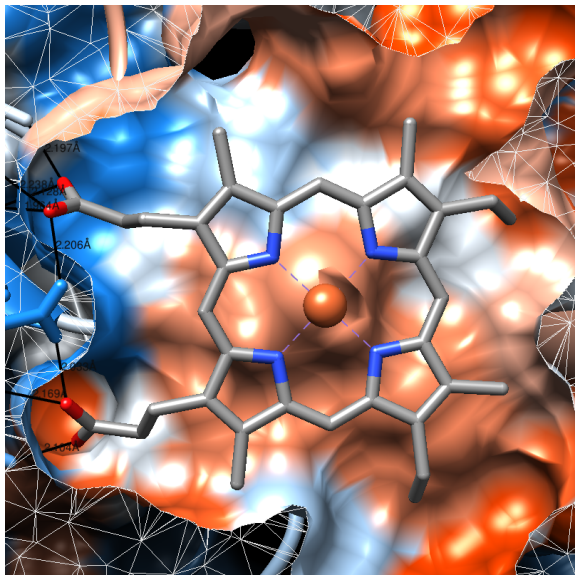


Figure 1.2.4: Protoporphyrin IX pseudobonds with Pn2.84.

1.2.2 Cinnamic acid

With the Pn2.84 protein and the best docked pose of protoporphyrin IX with Fe as a complex, molecular docking using Vina was carried out for trans-cinnamic acid resulting in nine different conformations, these results are shown in Table 3 The docked positions of the trans-cinnamic acid in the receptor structure interacted with different aminoacid residues of the protein (Fig. 1.2.5). Some of these aminoacids were found in the conserved domain for the active site of the enzyme, like Phe 107, Val 118, Phe 119, Arg 213, Ser 214, Gln 218, Val 305, Ala 306, Ile 371 and Val 375; the rest of the aminoacids were not marked as conserved in the active site domain but were present in the pockets found by CASTp3: Thr 310, Trp 313, Leu 374, Pro 376 and Phe 488.

Table 3: Docking results of cinnamic acid with protoporphyrin IX-Pn2.86 complex

Mode	Affinity (kcal/mol)	Dist. from best mode RMSD l.b	RMSD u.b
1	-6.66	0	0
2	-6.359	2.652	3.093
3	-6.147	1.105	1.982
4	-5.586	2.381	2.985
5	-5.577	3.276	5.403
6	-5.57	3.049	3.496
7	-5.414	5.377	5.865
8	-5.291	2.987	3.666
9	-5.271	4.219	5.511

A search for hydrogen bonds was performed with the best conformation (-6.66 kcal/mol) using UCSF ChimeraX allowing relax constraints of 0.4 Å and 20 degrees.[24;11;21] Four pseudobonds were found between the ligand and four aminoacids of the protein: two pseudobonds with Arg 213 with distances 2.378 Å and 2.539 Å; one pesudobond with Ser 214 of 2.329 Å; and one with Gln 218 of 2.312 Å. In Fig. 1.2.6 this pseudobonds and the ligand in its pocket are shown, also the distance between the trans-cinnamate and the Fe atom of the heme group is represented; the surface of the protein is colored based on its hydrophobicity using Kyte-Doolittle scale from lower (blue) to higher hydrophobicity (red). Reliability of the docking is supported by the fact that strong interactions between cinnamic acid and aminoacid residues Arg 213, Ser 214 and Gln 218 were found, consistent with what is found in the template crystal structure and the conserved domain for the active site. The interaction found between the cinnamic acid and the heme group and the slightly smaller distance between the cinnamic acid and the Fe atom found in this analysis compared to Zhang et al. 2020 (4.142 Å vs 4.4 Å), provide also a strong support for this docking methodology.

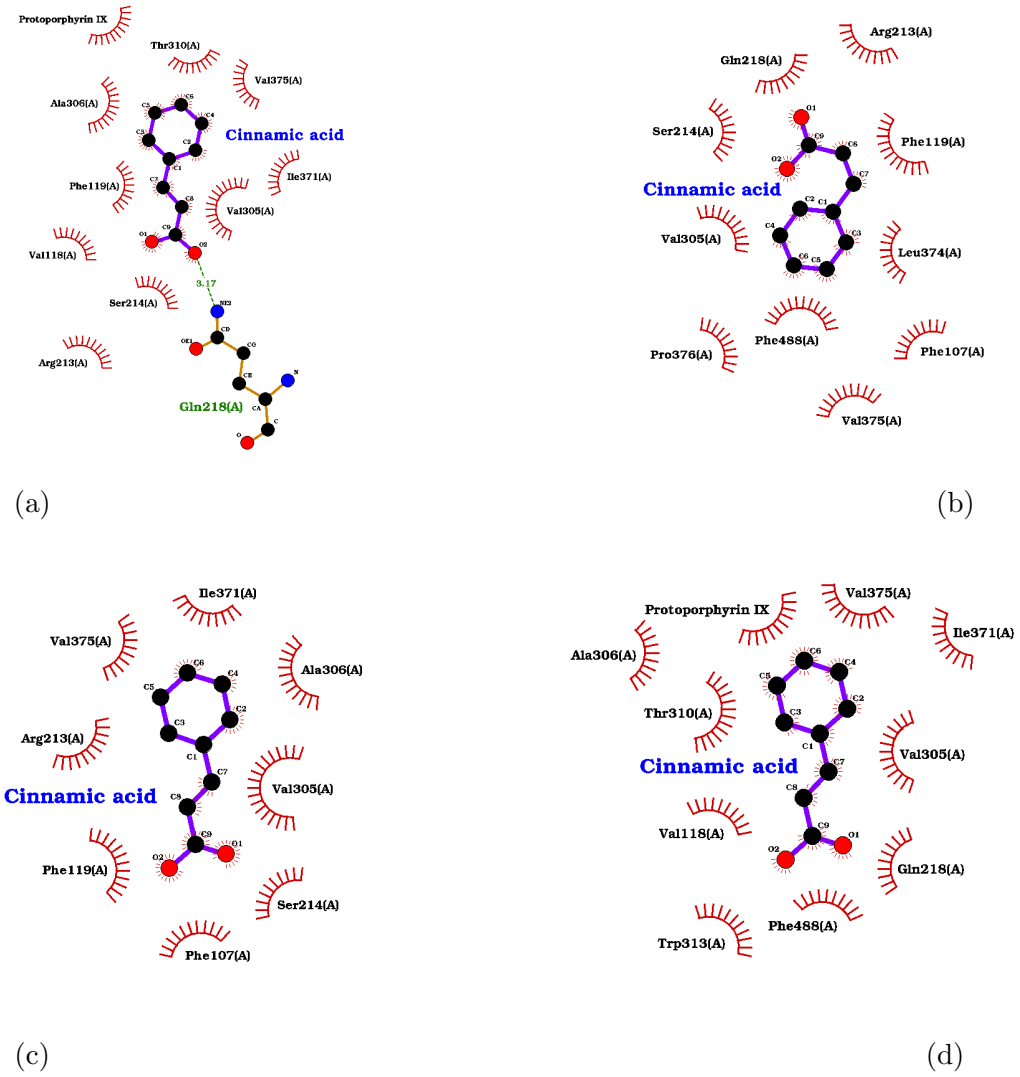


Figure 1.2.5: Interactions of cinnamic acid with protoporphyrin IX-Pn2.86 complex. Best four poses are shown in order a-d.

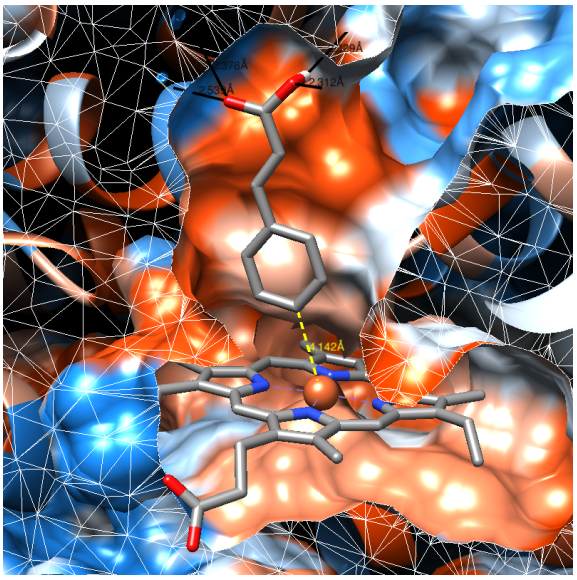
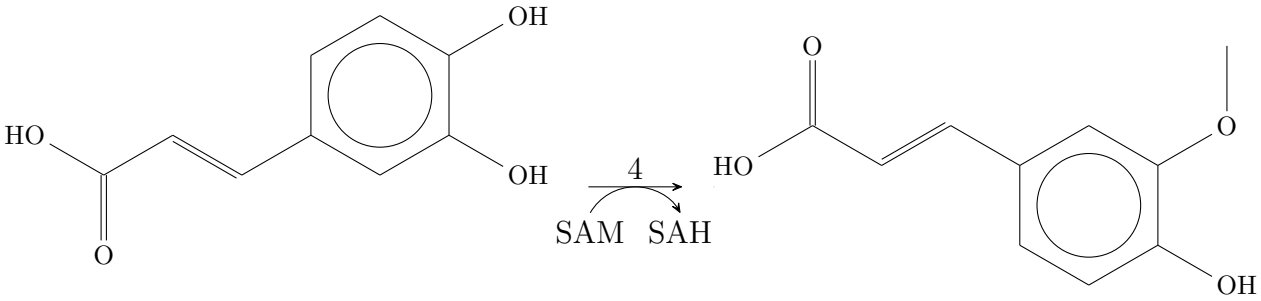


Figure 1.2.6: Cinnamic acid pseudobonds with protoporphyrin IX-Pn2.86 complex.

1.3 Pn1.1317

The fourth reaction of the pathway (Eq. 3) is catalyzed by the protein caffeic acid 3-O-methyltransferase, this enzyme is also part of the phenylpropanoid biosynthesis pathway. The gene Pn1.1317 was selected as the best candidate for enzyme 4 based on the fact that it obtained the highest score (651.5) over the threshold (499.33) for KEGG Orthology ID K13066, with E-value = 2.9×10^{-196}. This annotation was made with KofamKOALA.[4] With the exon usage analysis no relevant results were found.



(3)

Caffeic acid

Ferulic acid

The gene contains four exons that were considered when modeling the three-dimensional structure of the protein. The structure modeling was made by using an homology based methodology "SWISS-MODEL".[36] The selected template was an O-methyltransferase from *Fragaria ananassa* with PDB ID: 6I71; this protein has a 68.39% sequence identity to our query protein Pn1.1317 which makes it an appropiate template. The template was found with HHblits.[26] The model (Fig. 1.3.1) is found in a dimeric state, as the template is; it has a GMQE value of 0.84 and a QMEANDisCo Global value of 0.81 ± 0.05 (Fig. 1.3.2).[29] The enzyme requires a coenzyme SAM (S-adenosylmethionine) for its correct catalytic activity; this molecule is demethylated to SAH (S-adenosyl-L-homocysteine). The crystal structure of the template contained SAH binded to it, the biding site for SAH was conserved in Pn1.1317 so the model generated contained his molecule, the interactions between this ligand and the protein can be seen in Fig 1.3.3. The ligand 3D conformer structure for SAM and caffeic acid were obtained from PubChem with CIDs: 34755 and 689043.

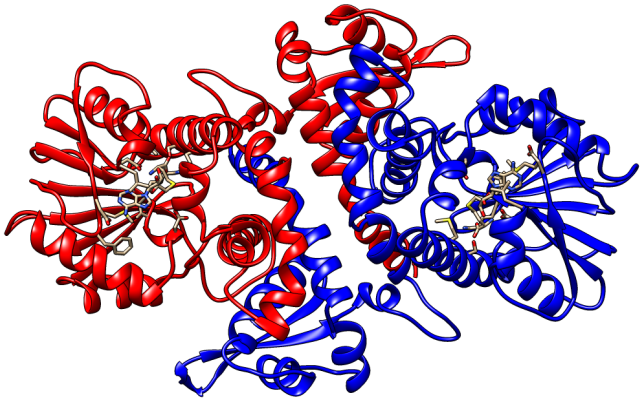


Figure 1.3.1: Pn2.84 three-dimensional model.

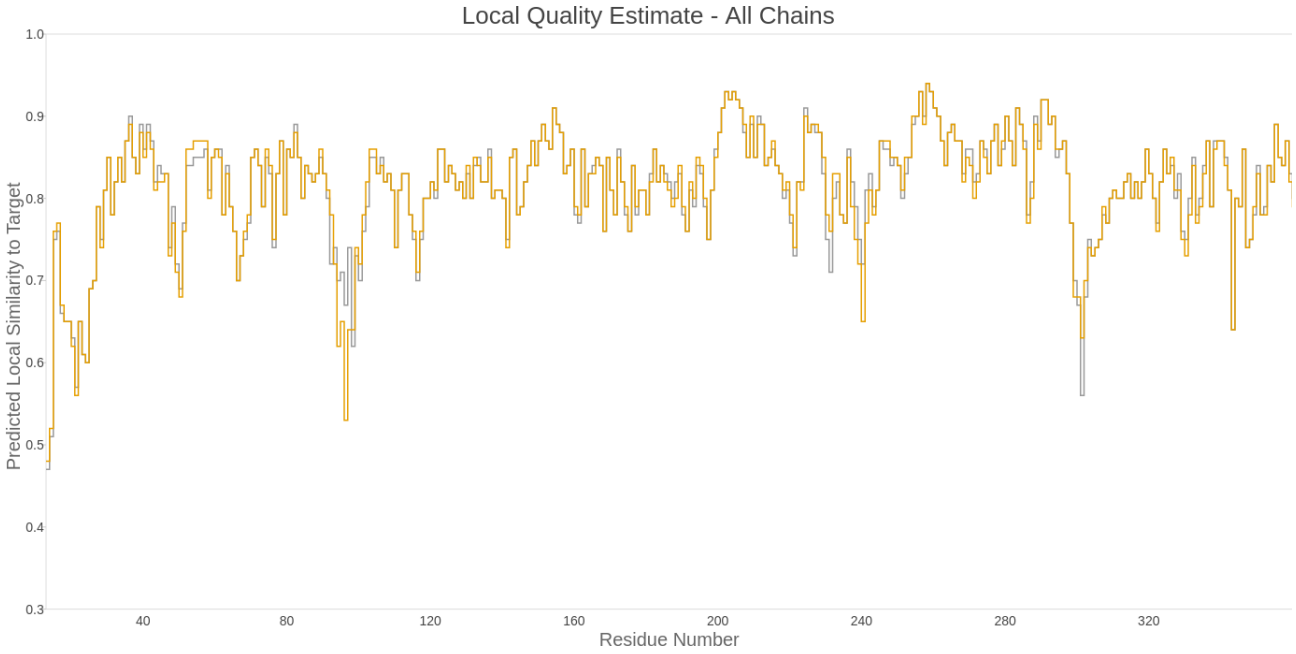


Figure 1.3.2: Pn1.1317 model QMEANDisCo.

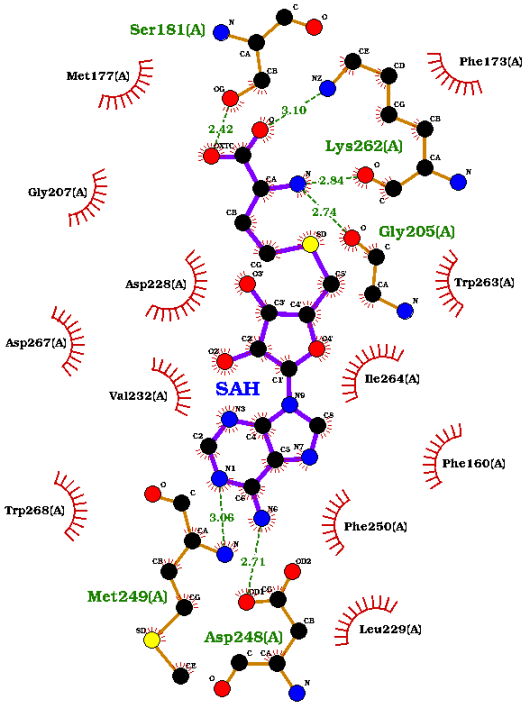


Figure 1.3.3: Interactions between SAH and Pn1.1317.

Before docking was performed the protein and the ligand were energy minimized. The protein structure was minimized using UCSF ChimeraX software[24;11] and the steepest descent algorithm for 1000 steps with a step size of 0.02 Å; the force field AMBER ff14SB was used for standar residues, for non standard residues the semi-empirical AM1-BCC model.[13;14;34] In the same way the ligands were energy minimized by using using the python library "rdkit" and the MMFF94 force field implemented within it.[16;32]

The preparation of the molecules for docking was made on AutoDockTools and Autodock Vina version 1.2.3 docked the protein and the ligand.[23;8;33] The grid box for the SAM docking was selected based on the conserved binding site for SAH found by homology modeling. In order to know the binding site for caffeic acid, blastp against PDB database was used to obtain similar protein sequences to Pn1.1317; 11 sequences with identity higher than 50% were selected.[1;2] Muscle aligned the sequences and based on the MSA (Multiple Sequence Alignment) a search for the best aminoacid substitution model was carried out using MEGA 11; the selected model was LG+G (lnL: -4126.121) this decision was supported by Bayesian Information Criterion (BIC: 8436.596) and Akaike's Information Criterion (8296.475).[9;30] With the selected model

and the MSA a Maximum Likelihood tree (ML) was build using MEGA 11, an heuristic search using Nearest Neighbour Interchange was performed and the branch support values were obtained from 100 Bootstrap pseudoreplicates.[30] Using the ML tree topology as a start point, a Bayesian Inference tree (BI) was build using MrBayes; two runs of four chains (1 cold chain and 3 heated chains) were run for one million generations and sampled every one thousand generations with a burn in of 25%, the amino acid substitution model was selected by sampling all of the fixed-rate models implemented in the program, in this way each model will contribute in proportion to its posterior probability.[28]

This analysis was run in the CIPRES Science Gateway portal.[20] To verify if convergence was reached, the results were analyzed with Tracer, the effective sample size for all the parameters was higher than 200: the log likelihood of the cold chain (LnL: 1721; Fig 1.3.4), the log likelihood of the prior (LnPr: 1668) and the total tree length (TL: 1658).[25] The resulting tree of the BI analysis can be seen in Fig. 1.3.5 Based on this results, the sequence with PDB ID: 1KYW, a Caffeic Acid/5-hydroxyferulic acid 3/5-O-methyltransferase from *Medicago sativa* was selected to define the binding site for caffeic acid. It is part of the sister group of Pn1.1317 and had the most significant alignment E-value (2×10^{-180}).

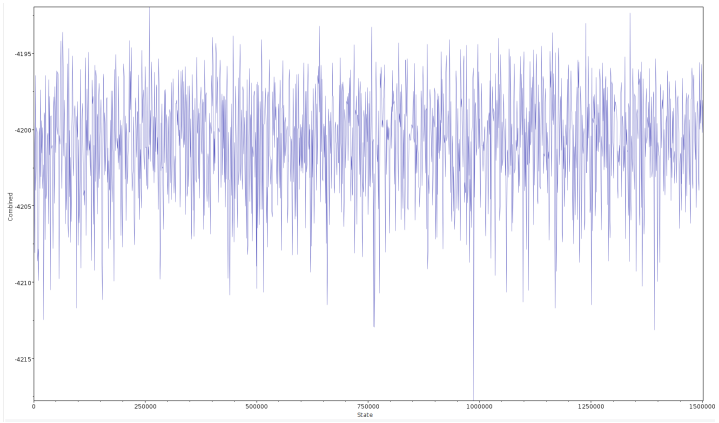


Figure 1.3.4: Convergence of log likelihood of the cold chain, BI analysis of Pn1.1317 phylogeny.

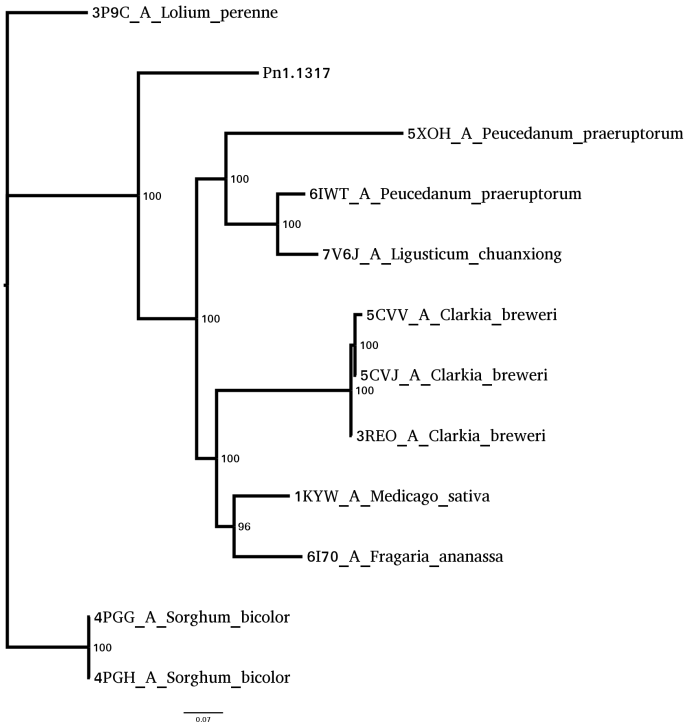


Figure 1.3.5: Phylogeny of Pn1.1317, Bayesian Inference.

Based on the results obtained for this protein by Zubieta et al. 2002, the interactions the substrate HFL (5-(3,3-dihydroxypropeny)-3-methoxy-benzene-1,2-diol) with 1KYW from the crystalographic structure (Fig. 1.3.6) and its homology with Pn1.1317 eleven aminoacid residues were selected as putative binding site for caffeic acid: Met 127, Leu 133, Ala 159, Phe 173, Met 177, Asp 267, Val 313, Ile 316, Met 317, His 320 and Asn 321.

CASTp3 found the protein's pockets.[31] To perform an adequate docking of both the SAM molecule and the caffeic acid, a sequential docking workflow was used as described in Vass et al. 2012 and in section 1.2.

1.3.1 SAM

The first molecule docked was the SAM molecule, as a result of molecular docking with Vina, nine different conformations of the molecule in the protein were found, these results are shown in Table 4. The docked positions of the protoporphyrin IX with Fe in the Pn2.84 structure interacted with different aminoacid residues of the protein (Fig. 1.3.7). Some of these aminoacids were found in the conserved domain for the SAM/SAH binding site of the enzyme, like Phe 160, Phe 173, Met 177, Ser 181, Gly 205, Gly 207, Asp 228, Leu 229, Val 232, Asp 248, Met 249, Lys 262, Trp 263, Ile 264, Asp 267 and Trp 268; the rest of the aminoacids were not marked as conserved in the heme binding site domain but were present in the pockets found by CASTp3: Asn 174, Ser 178, Gly 206, Ala 211, Phe 227 and Gly 247. In Fig. 1.3.8 the SAM in its pocket is shown, the surface of the protein is colored based on its hydrophobicity using Kyte-Doolittle scale from lower (blue) to higher hydrophobicity (red).

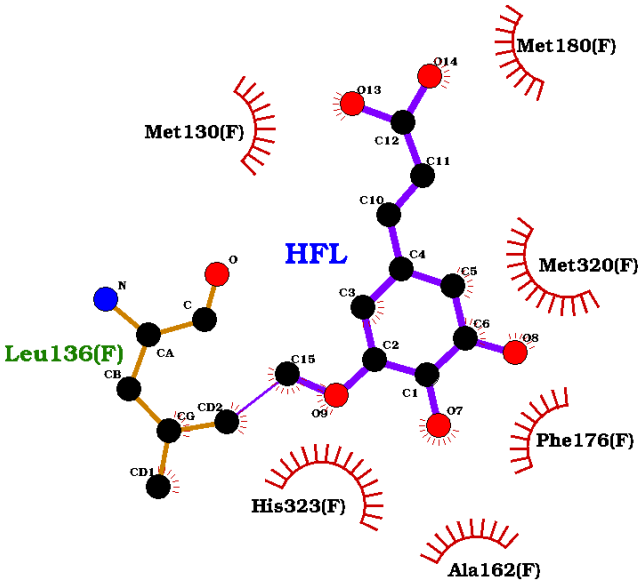


Figure 1.3.6: Interactions between caffeic acid and 1KYW.

Table 4: Docking results of SAM with Pn1.1317

Mode	Affinity (kcal/mol)	Dist. from best mode RMSD l.b	RMSD u.b
1	-8.594	0	0
2	-8.476	1.118	1.323
3	-8.364	2.652	8.627
4	-8.13	1.703	2.671
5	-8.113	2.585	8.519
6	-7.993	2.473	3.43
7	-7.778	2.394	3.229
8	-7.471	1.715	2.739
9	-7.239	3.722	8.452

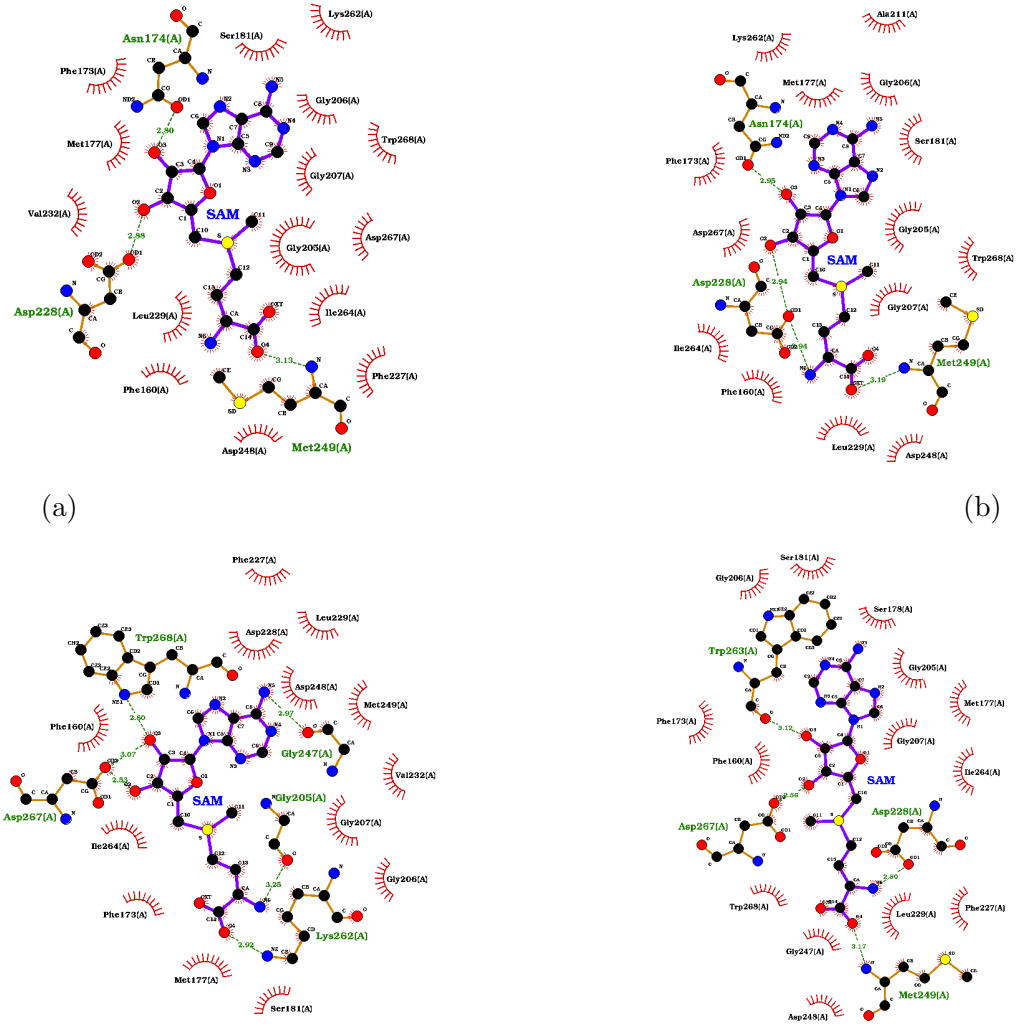


Figure 1.3.7: Interactions between SAM and Pn1.1317. Best four poses are shown in order a-d.

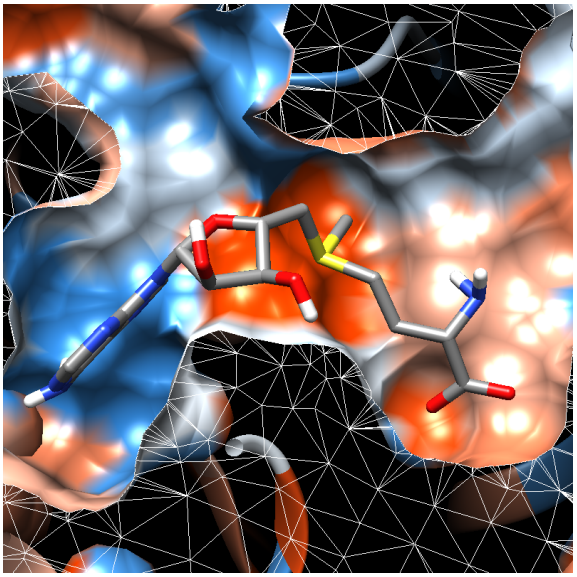


Figure 1.3.8: SAM in pocket of Pn1.1317.

1.3.2 Caffeic acid

With the Pn1.1317 protein and the best docked pose of SAM as a complex, molecular docking using Vina was carried out for caffeic acid resulting in nine different conformations, these results are shown in Table 5 The docked positions of the caffeic acid in the receptor structure interacted with different aminoacid residues of the protein (Fig. 1.3.9). Some of these aminoacids were found between the residues considered the active site of the enzyme and important for caffeic acid binding according to Zubieta et al. 2002: the hydrophobic residues Met 127, Phe 173, Met 177 and Met 317 sequester the phenyl ring that presents the reactive hydroxyl group to SAM, the residues Asp 267 and Asn 321 form hydrogen bonds with 5-OH group, interactions with the residues Asn-128, Met 127, Met 177, Val 313, and Ile 316 are necessary to bind the propanoid tail; the rest of the aminoacids were not marked as conserved in the active site domain but were present in the pockets found by CASTp3: Met 21, Leu 124, Trp 263 and His 266. In Fig. 1.3.10 the best docking pose of caffeic acid with SAM-Pn1.1317 in its pocket is shown, the surface of the protein is colored based on its hydrophobicity using Kyte-Doolittle scale from lower (blue) to higher hydrophobicity (red).

Table 5: Docking results of caffeic acid with SAM-Pn1.1317 complex

Mode	Affinity (kcal/mol)	Dist. from best mode RMSD l.b	RMSD u.b
1	-7.593	0.000	0.000
2	-7.298	0.769	1.874
3	-6.795	1.031	5.968
4	-6.067	1.058	5.939
5	-5.836	1.535	5.852
6	-5.691	2.752	5.559
7	-5.648	1.486	2.247
8	-5.393	2.771	5.849
9	-5.108	2.059	5.533

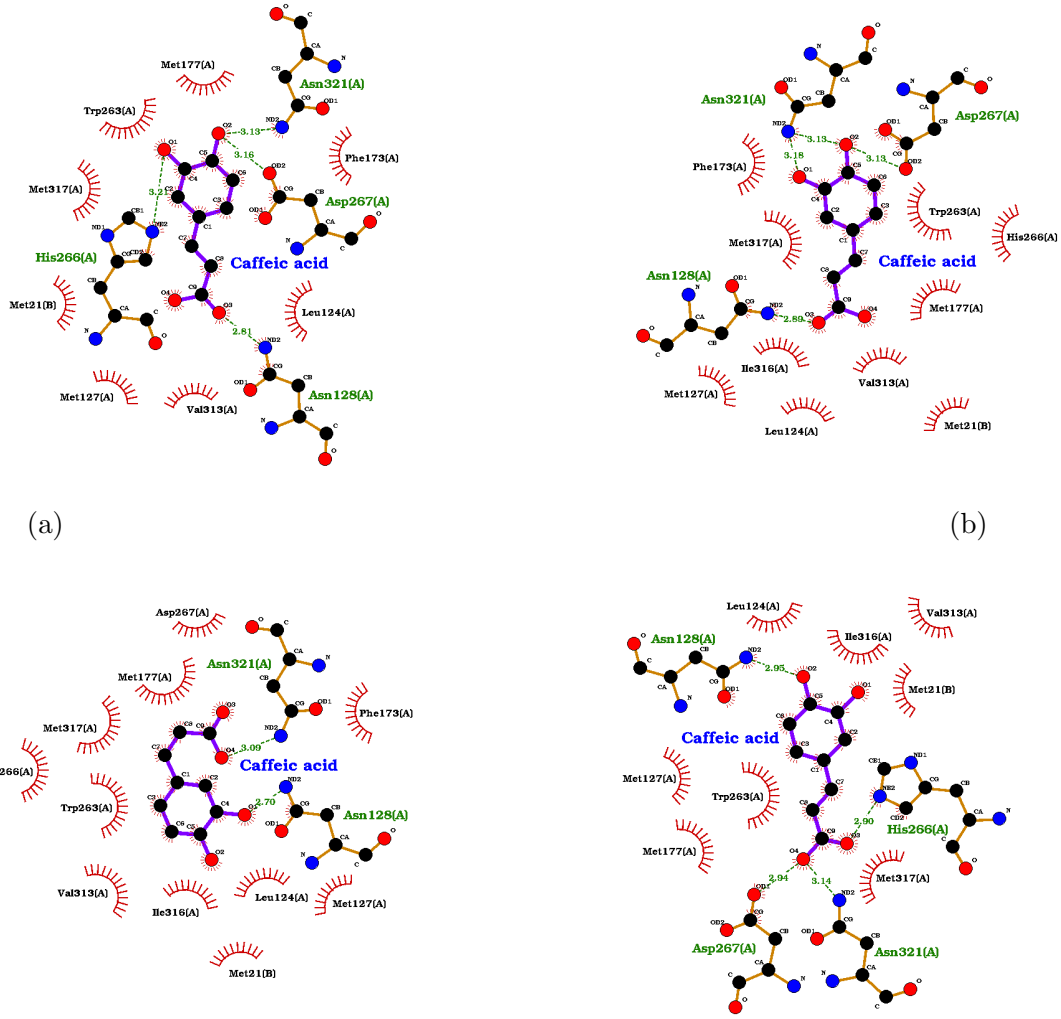


Figure 1.3.9: Interactions between SAM and Pn1.1317. Best four poses are shown in order a-d.

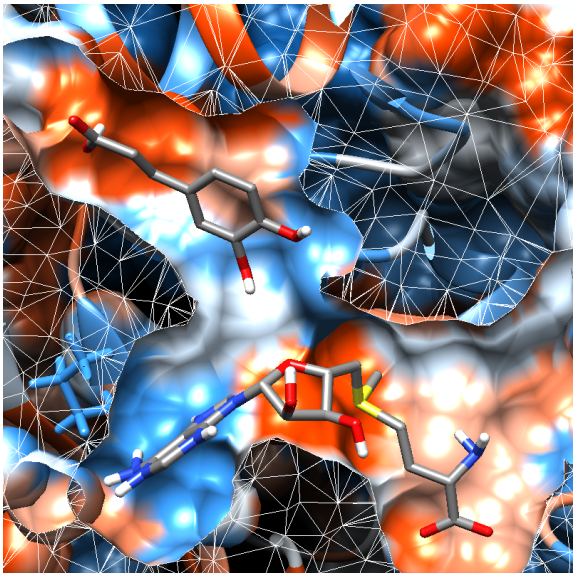
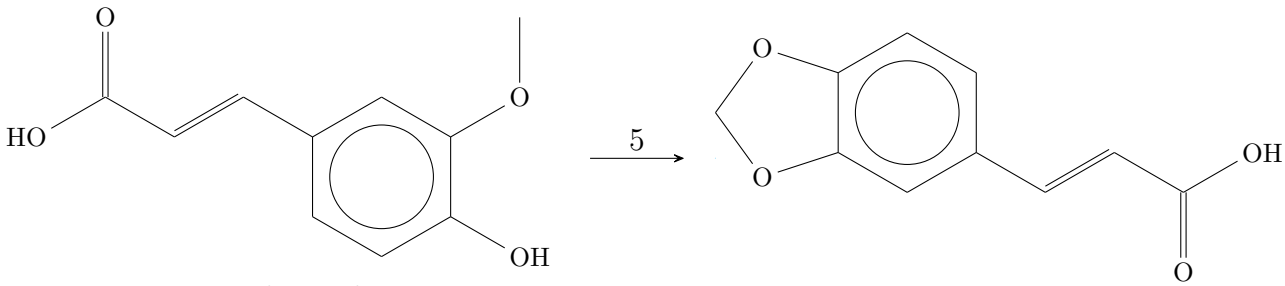


Figure 1.3.10: Caffeic acid in pocket of SAM-Pn1.1317 complex.

1.4 Pn3.4770

The fifth reaction of the pathway (Eq. 4) is catalyzed by a protein similar to CYP719A37. The gene Pn3.4770 was selected a candidate for enzyme 5 based on the fact that it has a 100% identity to CYP719A37 (E-value $\cong 0$) with accession number: QQS74306. The protein CYP719A37 has a length of 511 aminoacid residues, but the gene Pn3.4770 codifies for 578 aminoacids; this indicates that maybe there is an alternative splicing process affecting the expression of this gene. A Blast result and the exon usage analysis showed that from the three exons present in gene Pn3.4770, only the first two exons were expressed (Fig 1.4.2).



(4)

Ferulic acid

3,4-Methylenedioxy cinnamic acid

Only the first two exons of the gene were considered when modeling the three-dimensional structure of the protein. The structure modeling was made by using an artificial intelligence based methodology, the algorithm used for the 3D structure prediction of the protein was AlphaFold2 using the ColabFold resource.[15;22] ColabFold is a collaboratory space that allows users to perform three-dimensional structure prediction using AlphaFold2, AlphaFold2 multimer, MMseqs2 and HHsearch; this without having a super computer capable of running AlphaFold2 locally. The Local Distance Difference Test (lDDT) can be seen in Fig 1.4.3, this is a well suited score to assess local model quality.[19]

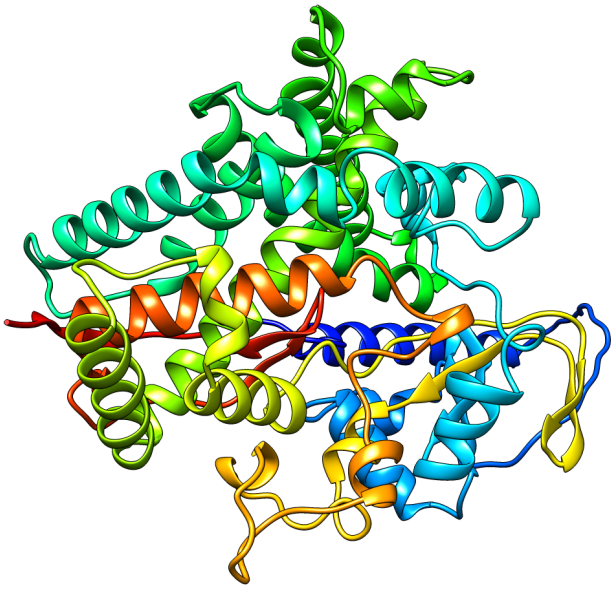


Figure 1.4.1: Pn3.4770 three-dimensional model.

The enzyme requires an heme molecule as a coenzyme for its correct catalytic activity; in this case the protoporphyrin IX containing Fe was selected based on the heme group present in the template for a "SWISS-MODEL" structure prediction that was made, the template was the CYP-450 17A1 protein from *Danio rerio* with PDB ID: 6b82; this model was not selected because of the low sequence identity with Pn3.4770 (24.62%). The ligand 3D conformer structure was obtained from PubChem with CID: 445858.

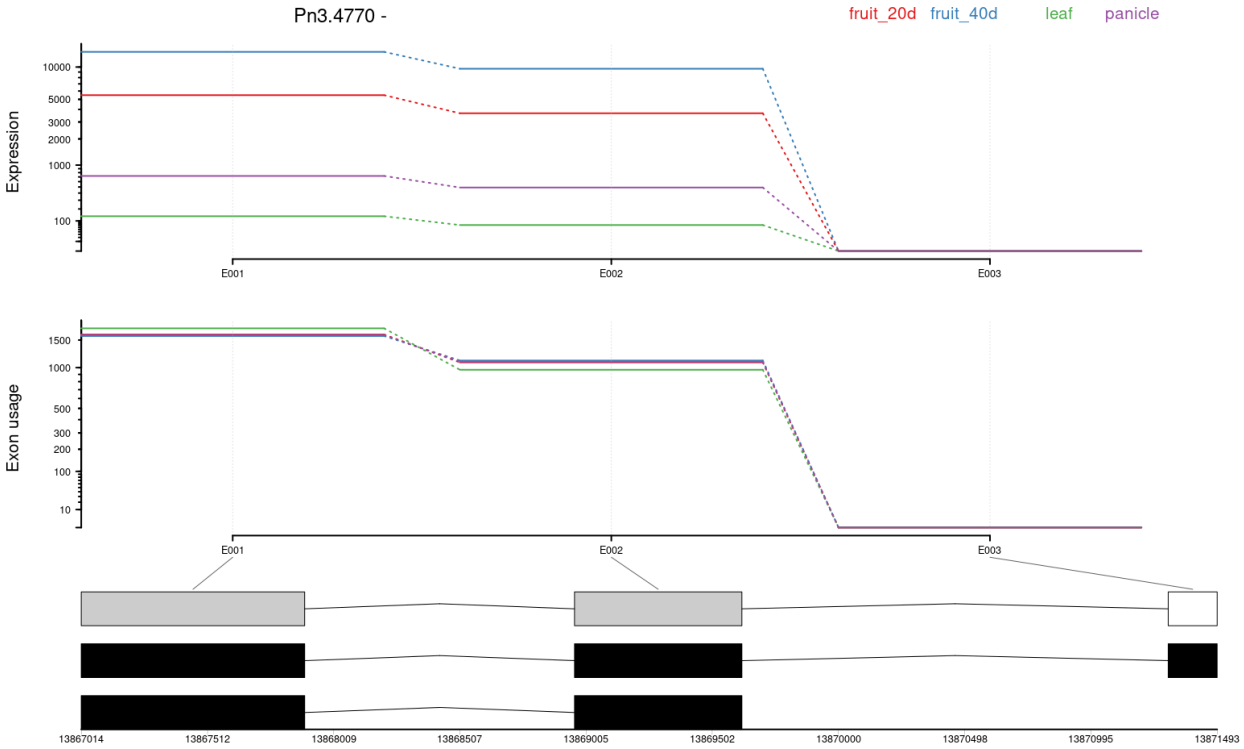


Figure 1.4.2: Exon usage analysis of Pn3.4770.

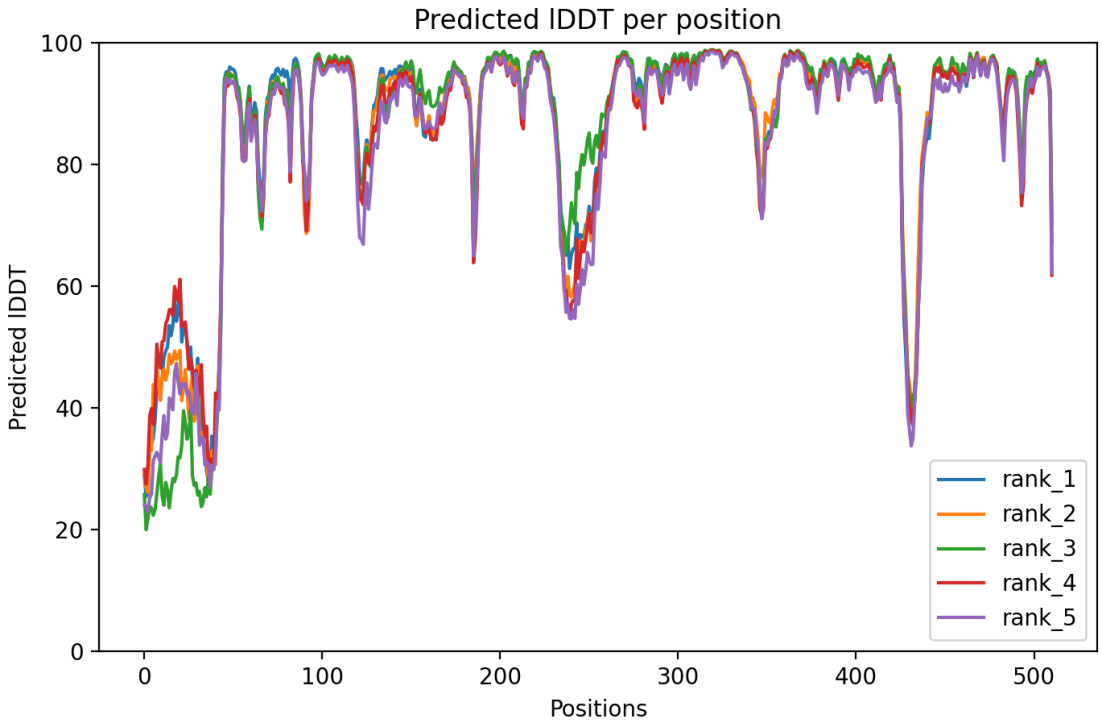


Figure 1.4.3: lDDT values of Pn3.4770.

Before docking was performed the protein and the ligand were energy minimized. The protein structure was minimized using UCSF ChimeraX software[24;11] and the steepest descent algorithm for 1000 steps with a step size of 0.02 Å; the force field AMBER ff14SB was used for standar residues, for non standard residues the semi-empirical AM1-BCC model.[13;14;34] In the same way the ligand was energy minimized by using using the python library "rdkit" and the MMFF94 force field implemented within it.[16;32]

The preparation of the molecules for docking was made on AutoDockTools and Autodock Vina version 1.2.3 docked the protein and the ligand.[23;8;33] The grid box for the docking was selected based on the conserved domains found with the NCBI conserved domain search interface; we found the heme binding site and putative chemical substrate binding pocket conserved domains for trans-cinnamate 4-hydroxylase from the NCBI-Curated Domains database; additionally, active site residues were selected from the template structure, which included mainly Arg 213, Ser 214 and Gln 218.[17;18]. CASTp3 found the protein's pockets.[31] To perform an adequate docking of both the heme group and the ferulic acid molecule, a sequential docking workflow was used as described in Vass et al. 2012 and in section 1.2.

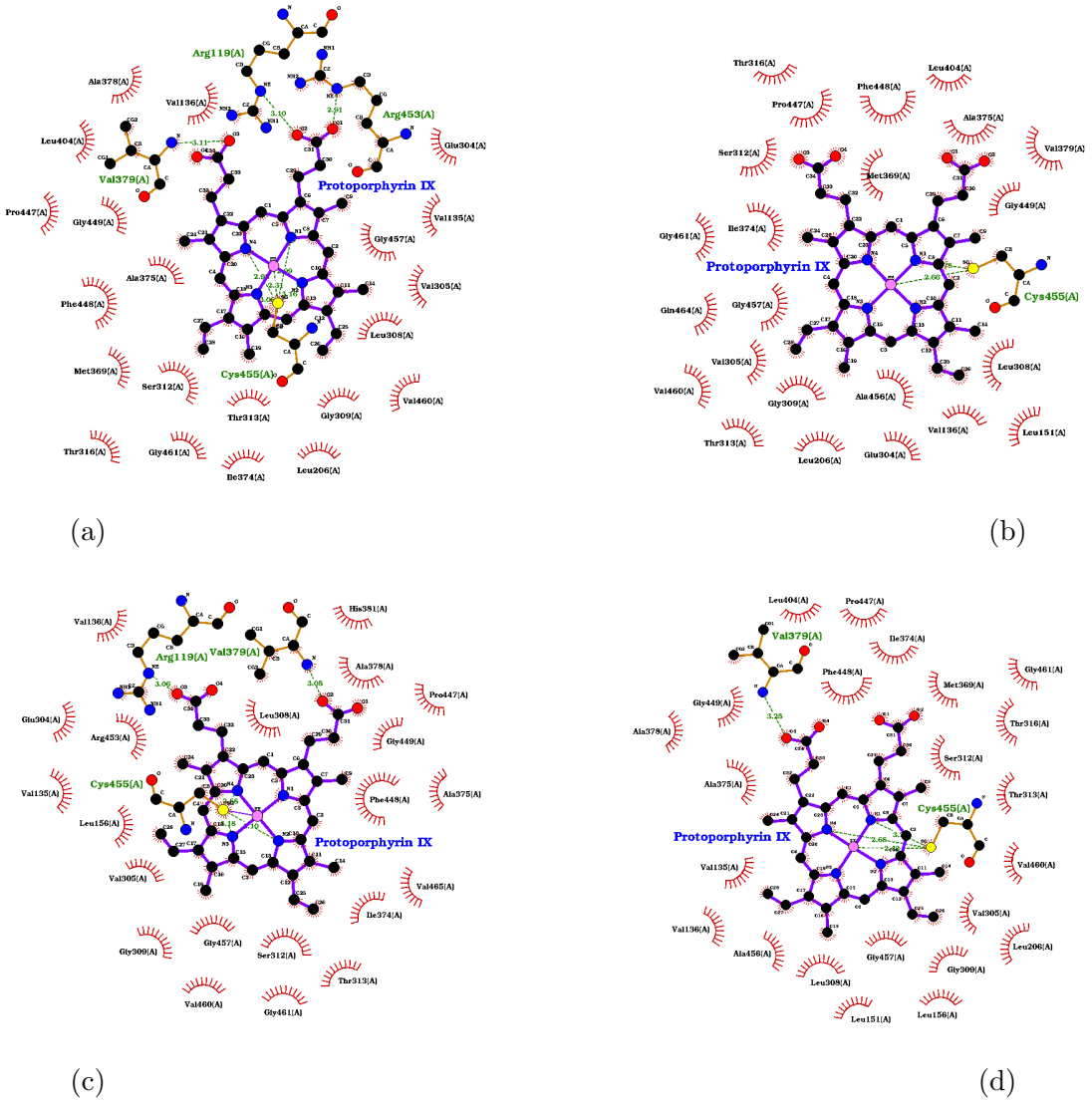
It is important to mention that the sequence used in this methodology (CYP719A37) has been isolated and its activity has been tested with different substrates. Its original substrate is feruperic acid, which has a two carbon extension in comparison with ferulic acid; when using ferulic acid as a substrate the expected product (3,4-methylenedioxy cinnamic acid) could not be detected. This indicates that this enzyme could not be able to catalyze the reaction, despite this, the docking was performed in order to compare the results of this enzyme with the one that is proposed in this work that could perform the reatcion.

1.4.1 Protoporphyrin IX with Fe

The first molecule docked was the protoporphyrin IX with Fe, as a result of molecular docking with Vina, nine different conformations of the molecule in the protein were found, these results are shown in Table 6. The docked positions of the protoporphyrin IX with Fe in the Pn3.4770 structure interacted with different aminoacid residues of the protein (Fig. 1.4.4). Some of these aminoacids were found in the conserved domain for the heme binding site of the enzyme, like Arg 119, Leu 156, Leu 206, Val 305, Leu 308, Gly 309, Ser 312, Thr 313, Ile 374, Ala 378, Val 379, His 381, Pro 447, Phe 448, Gly 449, Arg 453, Cys 455, Ala 456, Gly 457, Val 460 and Gly 461; the rest of the aminoacids were not marked as conserved in the heme binding site domain but were present in the pockets found by CASTp3: Val 135, Val 136, Leu 151, Glu 304, Thr 316, Met 369, Ala 375, Leu 404, Gln 464 and Val 465. In Fig. 1.4.5 the best docking pose of protoporphyrin IX with Fe with Pn3.4770 in its pocket is shown, the surface of the protein is colored based on its hydrophobicity using Kyte-Doolittle scale from lower (blue) to higher hydrophobicity (red).

Table 6: Docking results of protoporphyrin IX (Fe) with Pn3.4770

Mode	Affinity (kcal/mol)	Dist. from best mode RMSD l.b	RMSD u.b
1	-11.49	0	0
2	-11.04	2.227	6.083
3	-11.03	0.398	6.099
4	-10.48	2.164	6.827
5	-9.85	2.162	6.136
6	-9.165	3.37	7.65
7	-8.075	3.45	7.137
8	-7.811	2.199	6.704
9	-7.369	1.92	6.276



(a) (b)

(c) (d)

Figure 1.4.4: Interactions between protoporphyrin IX with Fe and Pn3.4770. Best four poses are shown in order a-d.

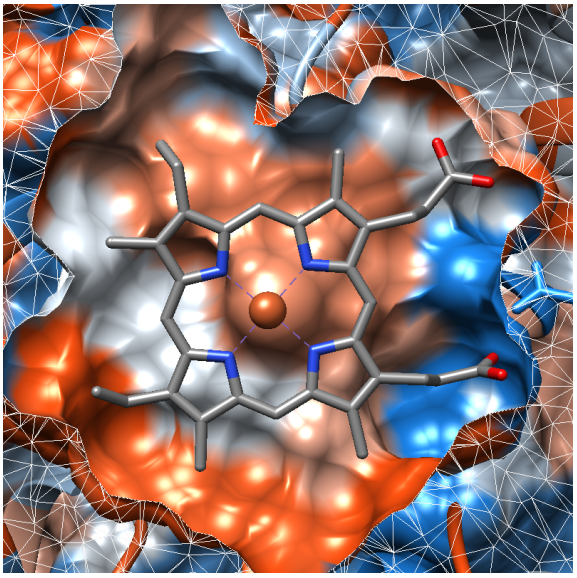


Figure 1.4.5: Protoporphyrin IX with Fe in pocket of Pn3.4770.

1.4.2 Ferulic acid

With the Pn3.4770 protein and the best docked pose of protoporphyrin IX with Fe as a complex, molecular docking using Vina was carried out for ferulic acid resulting in nine different conformations, these results are shown in Table 7 The docked positions of the trans-cinnamic acid in the receptor structure interacted with different aminoacid residues of the protein (Fig. 1.4.6). Some of these aminoacids were found in the conserved domain for the active site of the enzyme, like Leu 307, Leu 308 and Val 379; the rest of the aminoacids were not marked as conserved in the active site domain but were present in the pockets found by CASTp3: Leu 121, Phe 128, Asn 129, Val 136, Arg 239, Ser 312, Pro 380, Met 493 and Phe 495. In Fig. 1.4.7 the best docking pose of ferulic acid with protoporphyrin IX with Fe-Pn1.1317 complex in its pocket is shown, the surface of the protein is colored based on its hydrophobicity using Kyte-Doolittle scale from lower (blue) to higher hydrophobicity (red). The results show interactions between the ferulic acid and the Fe atom of protoporphyrin IX, but this interaction occurs in the popanoid tail instead of the 4-hydroxy-3-methoxyphenyl region.

Table 7: Docking results of ferulic acid with protoporphyrin IX-Pn3.4770 complex

Mode	Affinity (kcal/mol)	Dist. from best mode RMSD l.b	RMSD u.b
1	-6.166	0	0
2	-5.987	2.709	5.073
3	-5.888	3.896	5.423
4	-5.841	2.87	4.927
5	-5.739	2.885	6.541
6	-5.563	2.609	5.194
7	-5.491	3.6	5.481
8	-5.302	2.348	3.964
9	15.53	1.108	2.438

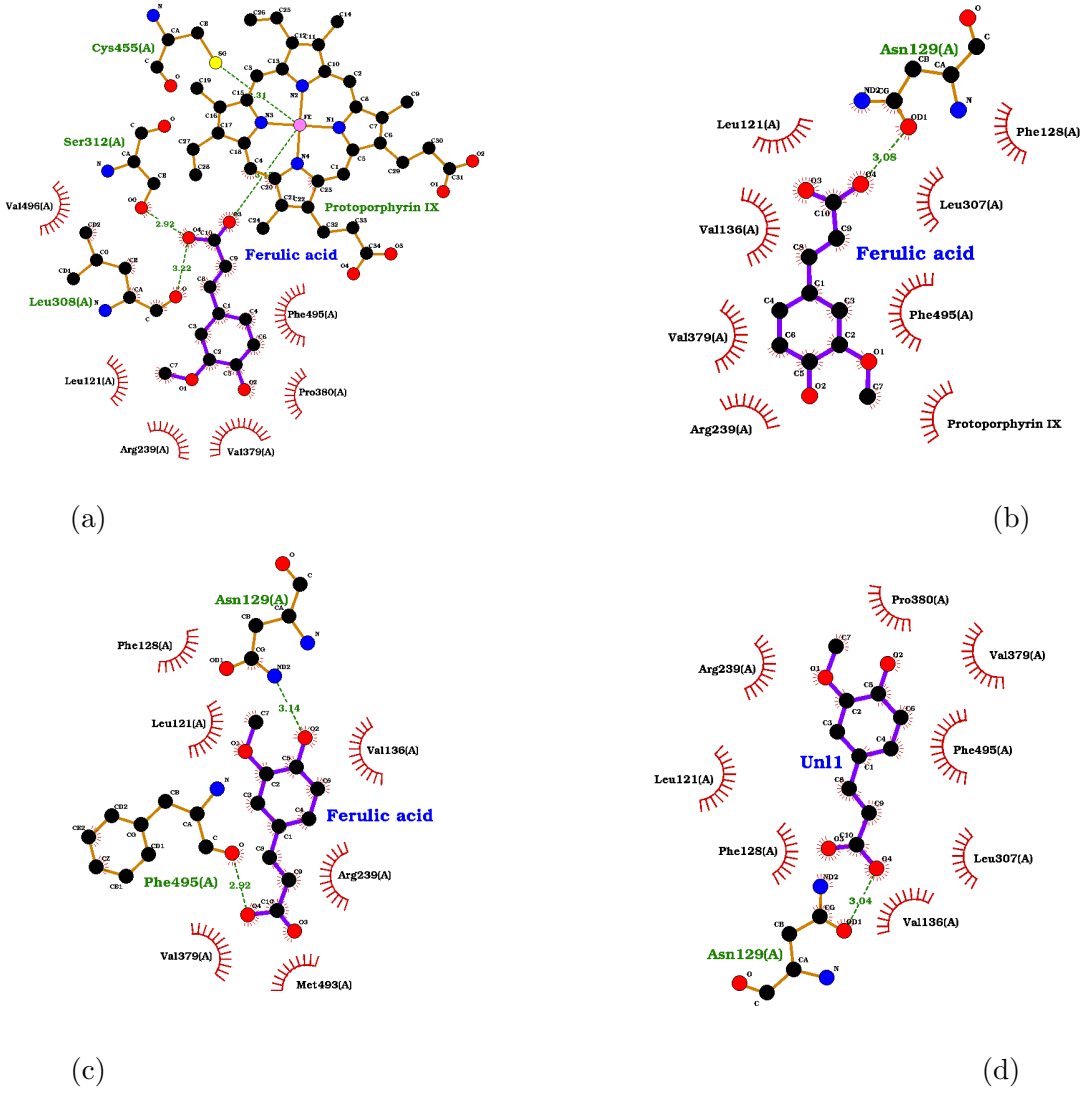


Figure 1.4.6: Interactions of ferulic acid with protoporphyrin IX-Pn3.4770 complex. Best four poses are shown in order a-d.

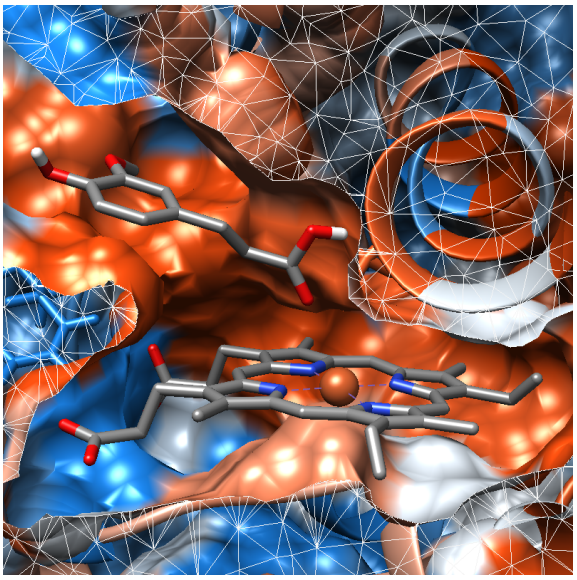


Figure 1.4.7: Ferulic acid in pocket of protoporphyrin IX-Pn3.4770 complex.

References

[1] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

[2] Altschul, S. F., Wootton, J. C., Gertz, E. M., Agarwala, R., Morgulis, A., Schäffer, A. A., and Yu, Y.-K. (2005). Protein database searches using compositionally adjusted substitution matrices. *The FEBS journal*, 272(20):5101–5109.

[3] Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from rna-seq data. *Nature Precedings*, pages 1–1.

[4] Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252.

[5] Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., and Schwede, T. (2017). Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. *Scientific reports*, 7(1):1–15.

[6] Community, T. G. (2022). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, 50(W1):W345.

[7] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21.

[8] Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. (2021). Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898.

[9] Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797.

[10] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):1–16.

[11] Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., and Ferrin, T. E. (2018). Ucsf chimerax: Meeting modern challenges in visualization and analysis. *Protein Science*, 27(1):14–25.

[12] Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., Wu, H., Qin, X., Yan, L., Tan, L., et al. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nature communications*, 10(1):1–11.

[13] Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. (2000). Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *Journal of computational chemistry*, 21(2):132–146.

[14] Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation. *Journal of computational chemistry*, 23(16):1623–1641.

[15] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

[16] Landrum, G. et al. (2013). Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*.

[17] Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Marchler, G. H., Song, J. S., et al. (2020). Cdd/sparcle: the conserved domain database in 2020. *Nucleic acids research*, 48(D1):D265–D268.

[18] Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., et al. (2015). Cdd: Ncbi's conserved domain database. *Nucleic acids research*, 43(D1):D222–D226.

[19] Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728.

[20] Miller, M. A., Pfeiffer, W., and Schwartz, T. (2011). The cipres science gateway: a community resource for phylogenetic analyses. In *Proceedings of the 2011 TeraGrid Conference: extreme digital discovery*, pages 1–8.

[21] Mills, J. and Dean, P. M. (1996). Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *Journal of computer-aided molecular design*, 10(6):607–622.

[22] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). Colabfold: making protein folding accessible to all. *Nature Methods*, pages 1–4.

[23] Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. (2009). Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16):2785–2791.

[24] Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., and Ferrin, T. E. (2021). Ucsf chimerax: Structure visualization for researchers, educators, and developers. *Protein Science*, 30(1):70–82.

[25] Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic biology*, 67(5):901–904.

[26] Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175.

[27] Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., and Huber, W. (2013). Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences*, 110(38):15377–15382.

[28] Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542.

[29] Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., and Schwede, T. (2020). Qmeandisco—distance constraints applied on model quality estimation. *Bioinformatics*, 36(6):1765–1771.

[30] Tamura, K., Stecher, G., and Kumar, S. (2021). Mega11: molecular evolutionary genetics analysis version 11. *Molecular biology and evolution*, 38(7):3022–3027.

[31] Tian, W., Chen, C., Lei, X., Zhao, J., and Liang, J. (2018). Castp 3.0: computed atlas of surface topography of proteins. *Nucleic acids research*, 46(W1):W363–W367.

[32] Tosco, P., Stiefl, N., and Landrum, G. (2014). Bringing the mmff force field to the rdkit: implementation and validation. *Journal of cheminformatics*, 6(1):1–4.

[33] Trott, O. and Olson, A. J. (2010). Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461.

[34] Tsai, K.-C., Wang, S.-H., Hsiao, N.-W., Li, M., and Wang, B. (2008). The effect of different electrostatic potentials on docking accuracy: a case study using dock5. 4. *Bioorganic & medicinal chemistry letters*, 18(12):3509–3512.

[35] Vass, M., Tarcsay, Á., and Keserű, G. M. (2012). Multiple ligand docking by glide: implications for virtual second-site screening. *Journal of computer-aided molecular design*, 26(7):821–834.

[36] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., et al. (2018). Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1):W296–W303.

[37] Zhang, B., Lewis, K. M., Abril, A., Davydov, D. R., Vermerris, W., Sattler, S. E., and Kang, C. (2020). Structure and function of the cytochrome p450 monooxygenase cinnamate 4-hydroxylase from sorghum bicolor. *Plant physiology*, 183(3):957–973.

[38] Zubieta, C., Kota, P., Ferrer, J.-L., Dixon, R. A., and Noel, J. P. (2002). Structural basis for the modulation of lignin monomer methylation by caffeic acid/5-hydroxyferulic acid 3/5-o-methyltransferase. *The Plant Cell*, 14(6):1265–1277.

Contents

List of Figures

List of Tables