Molecular modeling of Coq6, a ubiquinone biosynthesis flavin-dependent hydroxylase. Evidence of a substrate access channel

Alexandre Ismail

Université Pierre et Marie Curie

Ecole doctorale Complexité du Vivant (ED 515)

Thèse de Doctorat

Spécialité: Biochimie – Modélisation moléculaire

Molecular modeling of Coq6, a ubiquinone biosynthesis flavin-dependent hydroxylase

Evidence of a substrate access channel

Presented by

Alexandre ISMAIL

Directed by Caroline Mellot-Draznieks

Publicly presented on 5 January 2016

V15 – dernière version corrigée

Before a jury composed of

Olivier LEQUIN	Member / President
Sophie SACQUIN- MORA	Member / Reviewer
Bernard OFFMANN	Member / Reviewer
Catherine ETCHEBEST	Member
Vanessa PROUX	Member
Caroline MELLOT-DRAZNIEKS	Member
Marc FONTECAVE	Member

“To the everlasting glory of the Infantry – “
Frank Loesser

Acknowledgments

I would like to thank the following individuals and/or institutions for their help in the completion of this work:

- Dr. Vanessa Proux and Sup'Biotech
- Dr. Marc Fontecave
- Dr. Caroline Mellot-Draznieks
- Dr. Fabien Pierrel
- Dr. Murielle Lombard
- Dr. Lucie Gonzalez
- Dr. Catherine Etchebest
- Dr. Marc Gueroult
- Myriam Smadja
- My friends and family
- Jules Verne
- Michael Crichton
- Robert Heinlein
- Isaac Asimov

Molecular modeling of Coq6, a ubiquinone biosynthesis flavin-dependent hydroxylase

Evidence of a substrate access channel

Table of Contents

List of abbreviations

4-HB	4-hydroxybenzoic acid
4-HB6	3-hexaprenyl-4-hydroxybenzoic acid
4-HP	4-hydroxyphenol
4-HP6	3-hexaprenyl-4-hydroxyphenol
4-AB	4-aminobenzoic acid
4-AB6	3-hexaprenyl-4-aminobenzoic acid
4-AP	4-aminophenol
4-AP6	3-hexaprenyl-4-aminophenol
AMBER	Assisted Model Building and Energy Refinement
ATP	Adenosine triphosphate
ADP	Adenosine diphosphate
BLAST	Basic Local Alignment Search Tool
CATH	Class – Architecture – Topology – Homologous superfamily
CHARMM	Chemistry at Harvard Molecular Mechanics
CoQ	Coenzyme Q*, also known as ubiquinone*
CPMO	Cyclopentanone monooxygenase
DNA	Deoxyribonucleic acid
DSSP	Define Secondary Structure of Proteins
FAD	Flavin Adenine Dinucleotide
FPMO	Flavoprotein monooxygenase
GAFF	Generalized AMBER Force Field
GFP	Green Fluorescent Protein
GPU	Graphics Processing Unit
GR2	Glutathione Reductase - 2
GROMACS	Groningen Machine for Chemical Simulation
HMM	Hidden Markov Model
IMM	Inner Mitochondrial Membrane
IT	Information Technology
LINCS	Linear Constraint Solver
MAFFT	Multiple Alignment using Fast Fourier Transform
MD	Molecular Dynamics
MM	Molecular Mechanics
MSA	Multiple Sequence Alignment
NADH	Nicotinamide Adenine Dinucleotide
NADPH	Nicotinamide Adenine Dinucleotide Phosphate
NCBI	National Center for Biotechnology Information (USA)
NESG	Northeast Structural Genomics Consortium
NMR	Nuclear Magnetic Resonance
OPLS-AA	Optimized Potential for Liquid Simulations – All Atom
PAMO	Phenylacetone monooxygenase
PDB	Protein Data Bank
pHB	para-hydroxybenzoate
PHBH	para-hydroxybenzoate hydroxylase
PME	Particle Mesh Electrostatics
PSSM	Position Specific Scoring Matrix
Q	ubiquinone
QH	semiubiquinone

QH2	ubiquinol
QM	Quantum Mechanics
RESP	Restrained Electrostatic Potential
RMSD	Root Mean Square Deviation
ROS	Reactive Oxygen Species
SAM	S-adenosyl methionine
SCOP	Structural Classification of Proteins
VMD	Visual Molecular Dynamics
VDW	van der Waals

List of figures

Chapter 0

Chapter 1

Chapter 2

Chapter 3

Chapter 0 General introduction

1. General introduction

The present work is dedicated to the development of a molecular model of Coq6 in *S. cerevisiae*, an enzyme necessary for the biosynthesis of ubiquinone, a ubiquitous molecule in organisms that is essential for aerobic cellular metabolism.

Ubiquinone is a small molecule essential for electron and proton transfer among the protein complexes of the respiratory chain, which is responsible for generating the majority of cellular energy in oxygenic environments. These protein complexes are embedded in the inner mitochondrial membrane, and therefore ubiquinone has two distinct physico-chemical functional requirements: the ability to change redox state and lipid solubility. The redox ability to transfer electrons and protons is mediated by a fully substituted quinone ring, and lipid solubility is conferred by a polyprenyl tail.

The production of this essential molecule is effected by a dedicated aerobic ubiquinone biosynthesis pathway composed of over a dozen proteins[1] [2] [3] [4] as characterized in the model organism *S. cerevisiae*. There is evidence that these proteins work together in an obligate multi-protein complex, called the CoQ synthome[5] to transform a six-carbon ring (derived from chorismate, tyrosine, or 4-hydroxybenzoate[6]) into the fully substituted quinone ring of ubiquinone. The biosynthesis pathway as described in *S. cerevisiae* is resumed in **Figure 0.1** below.

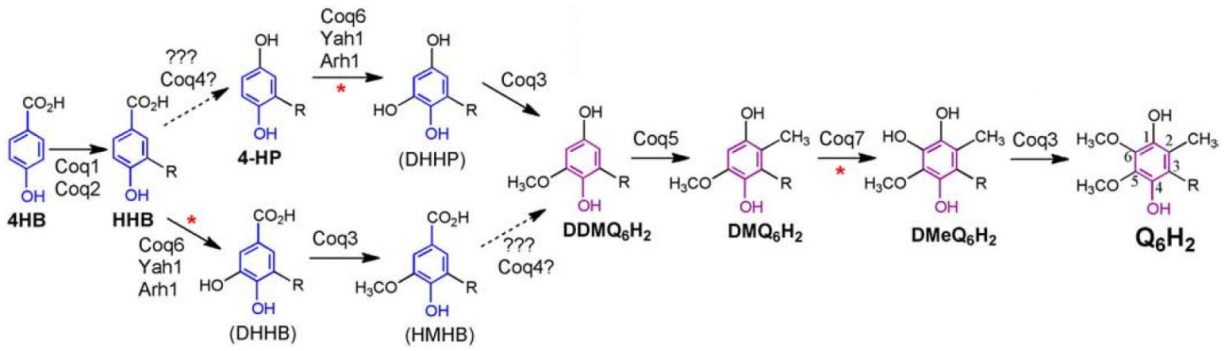

FIG 0.1: *The Q biosynthesis pathway from S. cerevisiae, adapted from "Coenzyme Q supplementation or over-expression of the yeast Coq8 putative kinase stabilizes multi-subunit Coq polypeptide complexes in yeast coq null mutants" (Biochimica and Biophysica Acta 2014).*[5]

However, the functional and structural interdependence[7] of the yeast Coq proteins makes the attribution of specific biosynthetic intermediates to specific enzymes difficult. Single Coq gene knockouts accumulate only early pathway intermediates,[8] and the precise ordering of reactions in the pathway is still ambiguous. One substitution, the addition of a methoxy group at the C5 ring position is partially understood as being catalyzed in two steps: a C5-hydroxylation followed by an O-methylation.[9]

Genetic and biochemical studies by our group have shown that the C5 hydroxylation is performed by the Coq6 enzyme,[10] and that this enzyme is likely to be an FAD dependent monooxygenase.[11] In this work, we present a further structure-function characterization of Coq6. Our group has isolated and purified Coq6 for biochemical study, but its structure has not yet been solved experimentally. Here, we establish the structure-function characterization on the basis of molecular modeling, site-directed mutagenesis, *in vivo* activity assays, and *in vitro* studies.

This structure-function characterization seeks to answer several specific questions:

1. What is the atomic-resolution structure of Coq6?
2. How does Coq6 bind its cofactor?
3. How does Coq6 bind its substrate?

The structural basis of this work is the creation of a homology model of the Coq6 enzyme. This model will then be used in the formulation of specific structure-function hypotheses which can be tested by functional *in vivo* assays of site-directed mutants informed by the homology model. The molecular modeling is performed by the author under the guidance of Caroline Mellot-Draznieks (Ph.D) at the Collège de France. The *in vitro* study of Coq6 was performed by Lucie Gonzalez (a Ph.D candidate) and Murielle Lombard (Ph.D), also at the Collège. The *in vivo* study of Coq6 was performed by Fabien Pierrel (Ph.D) at the University of Grenoble.

2. Modeling strategy

The modeling strategy consists of four main parts. First, we will create a panel of homology models of the Coq6 enzyme using several template coordinate sources and construction methods. Second, we will analyze the model structures for substrate binding regions and study their behavior over the course of molecular dynamics simulations. Third, we will dock substrate models into the substrate binding regions identified on the panel of homology models. This docking will be used as an *in silico* functional assay to identify a single substrate binding region and formulate structure-function hypotheses for it. This step will also be used to select a single model to retain for the next step. Fourth, we will use the selected model and the predicted enzyme-substrate interactions to rationally design mutants to test the structure-function hypotheses formulated in step three. Finally, the rationally designed mutations will be experimentally tested in a *in vivo* assay in *S. cerevisiae*.

3. Document structure

In this work we will describe these steps in more detail.

Chapter 1 will give detailed background information on known information on the ubiquinone biosynthesis system.

Chapter 2 will describe the computational strategy and methods we will use from a theoretical perspective. We will first translate the three main questions of the present study into general molecular

modeling tasks. These tasks are then further translated into specific techniques of molecular modeling. We will then compose these techniques into a larger strategy for modeling Coq6. Given the low sequence identity of Coq6 to the possible templates, the general goal of the strategy is to generate and test several possible models of Coq6 before simulating their interaction with a model substrate.

Chapter 3 will describe the application of the strategy developed in Chapter 2. We will first describe the construction of a set of Coq6 homology models using three independent methods. We will then test their structural stability using molecular dynamics.

Chapter 4 will describe the selection of Coq6 models based on attempts at docking a model substrate into the Coq6 active site. First, we will analyze our homology models to identify the active site and any possible substrate binding sites to derive geometric descriptors with which we will analyze the molecular dynamics trajectories created in Chapter 3.

Chapter 5 will describe the *in silico* testing of the substrate access channel in a yeast Coq6 mutant modeled from human clinical literature. We will further test the putative substrate access channel by using our homology model to rationally design additional mutations to block the channel. Finally, these predictions will be tested by *in vivo* activity assays.

Finally, in Chapter 6 we will present the conclusions of the current work and perspectives on the application of molecular modeling to the study of Coq6 and the entire Q biosynthesis pathway.

References

[1] Tzagoloff, A., & Dieckmann, C. L. (1990). PET genes of Saccharomyces cerevisiae. *Microbiological Reviews*, *54*(3), 211–225.

[2] Cui, Tie-Zhong, and Makoto Kawamukai. "Coq10, a Mitochondrial Coenzyme Q Binding Protein, Is Required for Proper Respiration in Schizosaccharomyces Pombe." *FEBS Journal* 276, no. 3 (February 1, 2009): 748–59. doi:10.1111/j.1742-4658.2008.06821.x.

[3] Pierrel, Fabien, Olivier Hamelin, Thierry Douki, Sylvie Kieffer-Jaquinod, Ulrich Mühlenhoff, Mohammad Ozeir, Roland Lill, and Marc Fontecave. "Involvement of Mitochondrial Ferredoxin and Para-Aminobenzoic Acid in Yeast Coenzyme Q Biosynthesis." *Chemistry & Biology* 17, no. 5 (May 28, 2010): 449–59. doi:10.1016/j.chembiol.2010.03.014.

[4] Allan, Christopher M., Agape M. Awad, Jarrett S. Johnson, Dyna I. Shirasaki, Charles Wang, Crysten E. Blaby-Haas, Sabeeha S. Merchant, Joseph A. Loo, and Catherine F. Clarke. "Identification of Coq11, a New Coenzyme Q Biosynthetic Protein in the CoQ-Synthome in Saccharomyces Cerevisiae." *Journal of Biological Chemistry*, January 28, 2015, jbc.M114.633131. doi:10.1074/jbc.M114.633131.

[5] He, Cuiwen H., Letian X. Xie, Christopher M. Allan, UyenPhuong C. Tran, and Catherine F. Clarke. "Coenzyme Q Supplementation or over-Expression of the Yeast Coq8 Putative Kinase Stabilizes Multi-Subunit Coq Polypeptide Complexes in Yeast Coq Null Mutants." *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1841, no. 4 (April 2014): 630–44. doi:10.1016/j.bbalip.2013.12.017.

[6] Meganathan, R. "Ubiquinone Biosynthesis in Microorganisms." *FEMS Microbiology Letters* 203, no. 2 (September 1, 2001): 131–39. doi:10.1111/j.1574-6968.2001.tb10831.x.

[7] He, Cuiwen H., Letian X. Xie, Christopher M. Allan, UyenPhuong C. Tran, and Catherine F. Clarke. "Coenzyme Q Supplementation or over-Expression of the Yeast Coq8 Putative Kinase Stabilizes Multi-Subunit Coq Polypeptide Complexes in Yeast Coq Null Mutants." *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1841, no. 4 (April 2014): 630–44. doi:10.1016/j.bbalip.2013.12.017.

[8] Hsieh, Edward J., Peter Gin, Melissa Gulmezian, UyenPhuong C. Tran, Ryoichi Saiki, Beth N. Marbois, and Catherine F. Clarke. "Saccharomyces Cerevisiae Coq9 Polypeptide Is a Subunit of the Mitochondrial Coenzyme Q Biosynthetic Complex." *Archives of Biochemistry and Biophysics* 463, no. 1 (July 1, 2007): 19–26. doi:10.1016/j.abb.2007.02.016.

[9] Padilla, S., U. C. Tran, M. Jiménez-Hidalgo, J. M. López-Martín, A. Martín-Montalvo, C. F. Clarke, P. Navas, and C. Santos-Ocaña. "Hydroxylation of Demethoxy-Q6 Constitutes a Control Point in Yeast Coenzyme Q6 Biosynthesis." *Cellular and Molecular Life Sciences* 66, no. 1 (January 1, 2009): 173–86. doi:10.1007/s00018-008-8547-7.

[10] Ozeir, Mohammad, Ulrich Mühlenhoff, Holger Webert, Roland Lill, Marc Fontecave, and Fabien Pierrel. "Coenzyme Q Biosynthesis: Coq6 Is Required for the C5-Hydroxylation Reaction and Substrate Analogs Rescue Coq6 Deficiency." *Chemistry & Biology* 18, no. 9 (September 23, 2011): 1134–42. doi:10.1016/j.chembiol.2011.07.008.

[11] Gin, Peter, Adam Y. Hsu, Steven C. Rothman, Tanya Jonassen, Peter T. Lee, Alexander Tzagoloff, and Catherine F. Clarke. "The Saccharomyces Cerevisiae COQ6 Gene Encodes a Mitochondrial Flavin-Dependent Monooxygenase Required for Coenzyme Q Biosynthesis." *Journal of Biological Chemistry* 278, no. 28 (July 11, 2003): 25308–16. doi:10.1074/jbc.M303234200.

Chapter 1 Introduction to ubiquinone biosynthesis

1. General introduction to ubiquinone and its role in cellular metabolism

The present work is dedicated to the development of a molecular model of Coq6 in *S. cerevisiae*, an enzyme necessary for the biosynthesis of ubiquinone, a ubiquitous molecule in organisms essential for aerobic cellular metabolism. After a preliminary presentation of the structure of ubiquinone, this chapter will introduce four key areas of background knowledge necessary to the development of the Coq6 molecular model. These are:

- Structures and functions of ubiquinone
- The ubiquinone biosynthesis pathway
- Current state of knowledge of Coq6
- The general structure of Q biosynthesis monooxygenases

We will first describe the structure and functions of ubiquinone. We will then review the existing knowledge of the ubiquinone biosynthesis pathway in our eukaryotic model organism *S. cerevisiae*. In order to introduce structural and functional nomenclature for the description of Coq6, we will review the known structure-function data for a related and extensively characterized enzyme, *para*-hydroxybenzoate hydroxylase (PHBH). Coq6 is predicted to use the same global fold and the same cofactor (flavin adenine dinucleotide, FAD) to perform the same reaction (nucleophilic hydroxylation) on a similar (para-hydroxybenzoate, pHB) substrate. We will conclude with a description and a discussion of known literature data for Coq6 and the formulation of specific questions to be addressed in this work.

1.1 What is ubiquinone?

Ubiquinone is a small molecule present in almost all cell membranes,[1] although it is found in highest concentrations in mitochondria.[2] Ubiquinone is a member of a more general family of molecules, the isoprenoid quinones, so named because they consist of a quinone moiety substituted with varying functional groups, one of which is an isoprenoid tail whose length varies between species. In *S. cerevisiae* it is six isoprene units long, whereas it is eight units long in *E. coli* and ten units long in *H. sapiens*. The isoprenoid tails give members of this family solubility in lipid membranes, which is important for their localization and function within the cell. The quinone moiety can be modified by different substituents which modulate the redox potentials of the compound to match their functional biochemical requirements, typically electron transfer in respiratory chains.[3] Classification according to the quinone moiety generates two main families: naphthoquinones (which have an additional aromatic ring fused onto the quinone ring at the C5 and C6 positions) and benzoquinones (which have discrete functional groups on the quinone ring). These structural features are summarized in **Figure 1.1**. Different types of quinones have different redox potentials, making each one suited for a specific set of biochemical reactions, sometimes under different redox conditions in different environments.

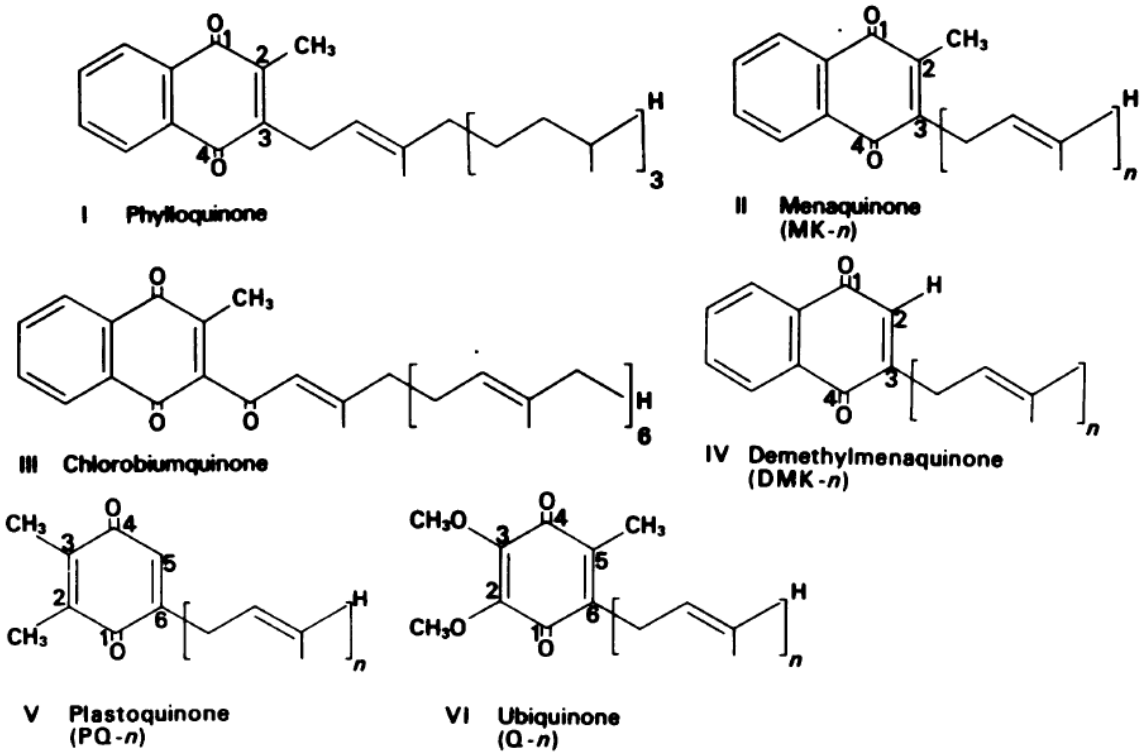

FIG 1.1: *Structures of the families of isoprenoid quinones. Inter-species complementation studies reveal that the precise aliphatic tail length is not essential to molecular function. However, the differing structures of the aromatic centers give the quinone variants different redox potentials. Adapted from* Distribution of isoprenoid quinone structural types in bacteria (Collins and Jones 1981).[3]

1.2 Structure of ubiquinone

Ubiquinone is a member of the benzoquinones, and it is the main electron transfer quinone of eukaryotes. Ubiquinone consists of a fully substituted quinone ring bearing a polyprenyl tail. The quinone ring bears two redox active hydroxyl groups, enabling ubiquinone to exist in three redox states: fully oxidized (as ubiquinone, hereafter referred to as Q), partially reduced (as semiubiquinone, hereafter referred to as QH), and fully reduced (as ubiquinol, hereafter referred to as QH_2). The polyprenyl tail confers lipid solubility, localizing the molecule to membranes. These two general physico-chemical properties enable Q to perform several essential functions in cellular metabolism. It can act as a lipid soluble electron transfer reagent in the electron transport chain, an anti-oxidant for membrane lipids, proteins, and DNA, and as a structural membrane lipid. These structures are resumed in **Figure 1.2** below.

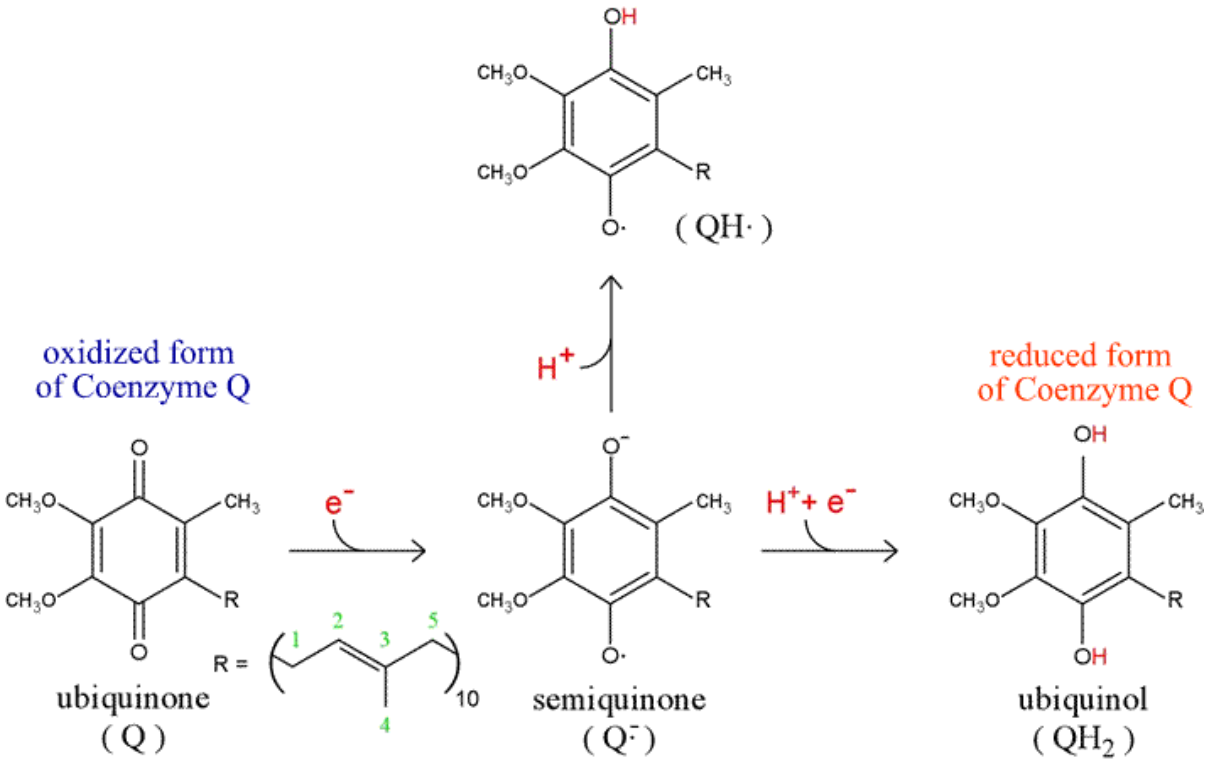

FIG 1.2: *Ubiquinone in its various redox states. The isoprenyl tail is indicated in parentheses. Oxidation and reduction occur at the two phenolic oxygens.*

1.3 Functions of ubiquinone

1.3.1 A lipid soluble redox agent in the electron transport chain

Q is perhaps best known for its essential role as an electron transfer reagent in the mitochondrial electron transport chain, where it is required to transfer electrons from Complex I and Complex II to Complex III.[4] These protein complexes are embedded in the inner mitochondrial membrane (IMM) as shown in **Figure 1.3**. Q's quinone ring is responsible for electron and proton transfer and the lipid tail makes the molecule strongly hydrophobic, localizing it inside the membrane. These properties make Q both essential to cellular function and difficult to transport through aqueous compartments, and it is poorly absorbed from the gastro-intestinal tract.[5] Therefore, all cells have an endogenous capacity to biosynthesize Q. It is not surprising that the site of highest Q utilization, the mitochondria, is also a major site of Q biosynthesis. Generally, the higher the metabolic requirements of a given cell type, the more Q it contains. There are two reasons for this: Q is required to generate adenosine triphosphate (ATP) by aerobic electron transfer, and it is also required to neutralize the reactive oxygen species (ROS) which are by-products of the same process. ROS can damage the electron transport chain and other cellular components through uncontrolled oxidation.

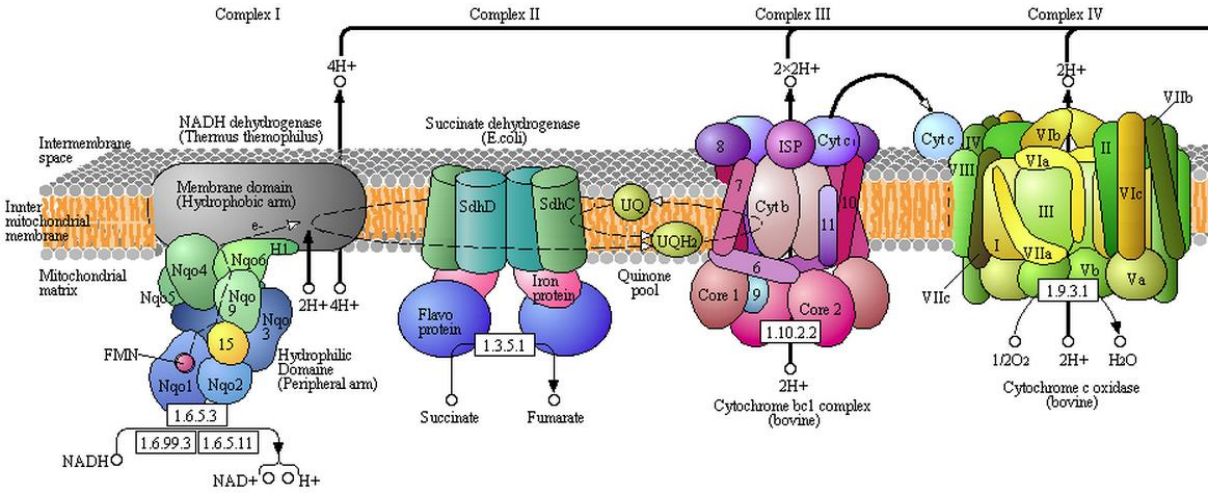

FIG 1.3: *Multi-protein complexes of the electron transport chain from KEGG, the Kyoto Encyclopedia of Genes and Genomes.*[6]

The biochemical goal of deriving cellular energy from ingested food through aerobic respiration is the transformation of redox potential from a variety of chemically *diverse* food molecules (carbohydrates, lipids, and proteins) into a single *common* energy transfer molecule: adenosine triphosphate (ATP). This transfer of redox potential is accomplished through the biochemical manipulation of some of the most generic constituents of matter: the proton and the electron.

This occurs in three main phases. In the first phase, food molecules of varying types are transformed into acetate for processing in the citric acid cycle, which produces a small set of more generic reduced carbon metabolites, namely NADH and succinate. In the second phase, the reductive potential of these specific molecules (NADH and succinate) is then converted into an even more generic physical form: a concentration gradient across a membrane. The redox potential of many individual NADH and succinate molecules is converted into a single common proton gradient on a supra-molecular scale: the inter-membrane space of the mitochondria. This gradient is formed and maintained by active proton transfer from the mitochondrial matrix to the inter-membrane space through the action of the electron transport chain. In the third phase, the energy stored in this concentration gradient is used for the synthesis of a specific redox transfer molecule, ATP.

The electron transport chain is a system of several separate multi-protein complexes that has evolved to use electron transfer among proteins *within* the inner mitochondrial membrane (IMM) to drive proton transfer *across* the inner mitochondrial membrane. Electrons are passed from NADH and succinate to species (small molecule redox agents like Q, or macromolecular redox agents like cytochrome C) having progressively higher oxidation potentials, with the free energy released used to translocate protons. This strongly implies the need for an electron and proton transfer reagent that can be localized within the inner mitochondrial membrane, and biochemistry has evolved to use and synthesize a specific molecule for this role: ubiquinone (Q). Here we will briefly summarize the role of Q as a lipid-soluble redox agent in the electron transport chain.

The electron transport chain is composed of four large protein complexes, named Complexes I-IV. Each complex has a different role, but they work together to use the reductive potential of specific metabolites (NADH and succinate) to translocate protons across the IMM.

Complex I takes electrons and protons from NADH to reduce Q to QH_2 and concomitantly translocate four protons from the matrix to the intermembrane space. Complex II takes electrons and protons from succinate to reduce Q to QH_2, but does not translocate protons. Together, Complexes I and II use Q as an electron and proton acceptor to create a "pool" of reduced QH_2 in the IMM, carrying the protons and electrons abstracted from food.

Complex III uses this pool of reduced QH_2 for two purposes. Protons are abstracted from QH_2 (regenerating Q) and translocated to the inter-membrane space (contributing to the proton gradient), while the electrons are transferred to another electron acceptor of higher oxidation potential, cytochrome C, to be passed on to Complex IV.

Complex IV will oxidize cytochrome C and use these electrons to drive one last proton translocation event. Four electrons from four cytochrome C moieties are transferred to a final acceptor: molecular oxygen. These four electrons must be paired with four protons, which are taken from the mitochondrial matrix, and molecular oxygen is finally reduced to water.

Ubiquinone is therefore essential for the proton translocation events catalyzed by Complexes I and III, without which the mitochondrial proton gradient could not be established and aerobic respiration could not occur.

1.3.2 An antioxidant for membrane lipids, proteins, and DNA

The ability to support multiple oxidation states, including the QH radical, enables this small molecule to act as a more general anti-oxidant, preventing peroxidation of membrane lipids, DNA, and proteins. This is an important function because Q is present in high quantities in cellular compartments with elevated redox activity, such as mitochondria and lysosomes,[7] as well as compartments which may require careful control of redox events, such as protein maturation in the Golgi vesicles.[5]

The inner mitochondrial membrane is an important source of ROS, primarily the superoxide anion, produced from the incomplete one-electron reduction of dioxygen and other unintended electron "leaks" from the electron transfer chain. Generation of ROS can be harmful to the cell because of their ability to initiate radical chain reactions, increasing the stoichiometry of damage far beyond the initial quantity of ROS produced. About 0.1-2% of the electrons processed by the electron transfer chain are lost to incomplete reduction of dioxygen, yielding ROS. Proximal targets of these ROS are the protein complexes of the electron transfer chain themselves, some of which contain iron-sulfur clusters. Oxidation of these proteins and their Fe-S clusters can release free iron into the cell, which is a potent source of the hydroxyl radical, which can initiate oxidation of lipids, proteins and DNA.

Q and QH can react with the superoxide anion to form hydrogen peroxide, which can then be detoxified by the action of catalase to yield water and oxygen.[8] By preventing the initiation steps of biomolecular

peroxidation, Q can greatly reduce the negative effects of ROS on the cell. In summary, Q participates in generating radical species as well as in neutralizing them; the control of this behavior depends on sub-cellular localization as well as the molecular environment and the presence of other redox active species.

1.3.3 A structural membrane lipid

Ubiquinone's isoprenyl tail also gives this molecule a non-redox role as a structural membrane lipid (similar to dolichol and cholesterol) where it appears to increase the stability of *E. coli* cell membranes in high-salt conditions[9] and reduce membrane permeability to protons and sodium.[10] The length of this polyprenyl chain varies among organisms, from six isoprene units in *S. cerevisiae*, to eight in *E. coli*, to ten in humans. Experimental evidence indicates that ubiquinones with tails of six or more isoprene units occupy a position between the membrane leaflets,[11] and that a tail of eight or more isoprene units may significantly enhance membrane stability.[12]

Neutron scattering experiments indicate that Q with six or more isoprene units assumes two main conformations in lipid bilayers.[13] In one, the polyprenyl tail and head occupy a position in between the lipid leaflets. In the other, the isoprenyl tail occupies largely the same position but the aromatic head is situated near the polar headgroups of the membrane lipids. This orientation places the bulk of the isoprenyl chain's surface area perpendicular to the membrane normal, which could make it an effective physical barrier to the uncontrolled permeation of protons and ions. This orientation is depicted schematically below in **Figure 1.4**.

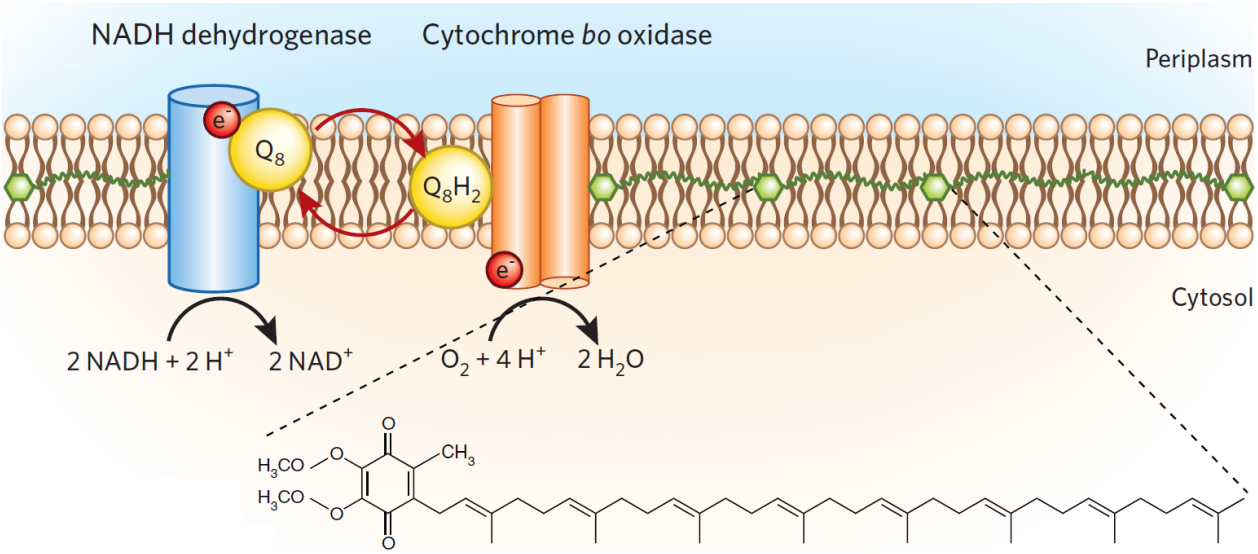

FIG 1.4: *Q in the membrane midplane. Adapted from* Is CoQ a membrane stabilizer? *(Nature Chemical Biology 2014).*[9]

2. The ubiquinone biosynthesis pathway in *S. cerevisiae*

2.1 Overview

The Q biosynthesis pathway in *S. cerevisiae* is composed of at least twelve proteins: Coq1-9,[14] [15] Arh1, Yah1,[16] and Coq11.[17] There is evidence that these proteins work together in an obligate multi-protein complex, called the CoQ synthome[18] to transform a six-carbon ring (derived from chorismate or tyrosine, both of which lead to 4-hydroxybenzoate[19]) into the fully substituted quinone ring of ubiquinone. However, the attribution of enzymes to biosynthesis intermediates is still incomplete, and new genes involved in Q biosynthesis are still being discovered. The biosynthesis pathway as currently described in *S. cerevisiae* is resumed in the **Figure 1.5** below.[18]

Coq1 is responsible for the synthesis of the polyprenyl tail. Coq2-7 are responsible for modifying and adding substituents to the eventual quinone ring. Coq6, which is the focus of this thesis, is responsible for the hydroxylation of the aromatic ring at the C5 position.[20] Coq8 contains a kinase domain which may regulate synthome assembly through protein phosphorylation as well as two flanking domains which may serve as protein-protein interaction surfaces for synthome assembly. Coq9 is likely to be involved in deaminating substrates bearing an amino group at position C4. Coq11 is likely to be a chaperone, perhaps assisting in the transport of Q biosynthesis intermediates.[17] Arh1 (an adrenodoxin reductase) and Yah1 (an adrenodoxin homolog) are implicated as part of an electron transfer system supplying reducing equivalents to the biosynthesis enzymes.[20]

In *S. cerevisiae* the quinone ring is derived from either tyrosine or chorismate.[21] Both of these precursors are processed to yield the common product 4-hydroxybenzoate (4-HB), from which the atom numbering of the quinone ring is derived. We will introduce this numbering here (see **Figure 1.6**) as it will be used frequently in this work.

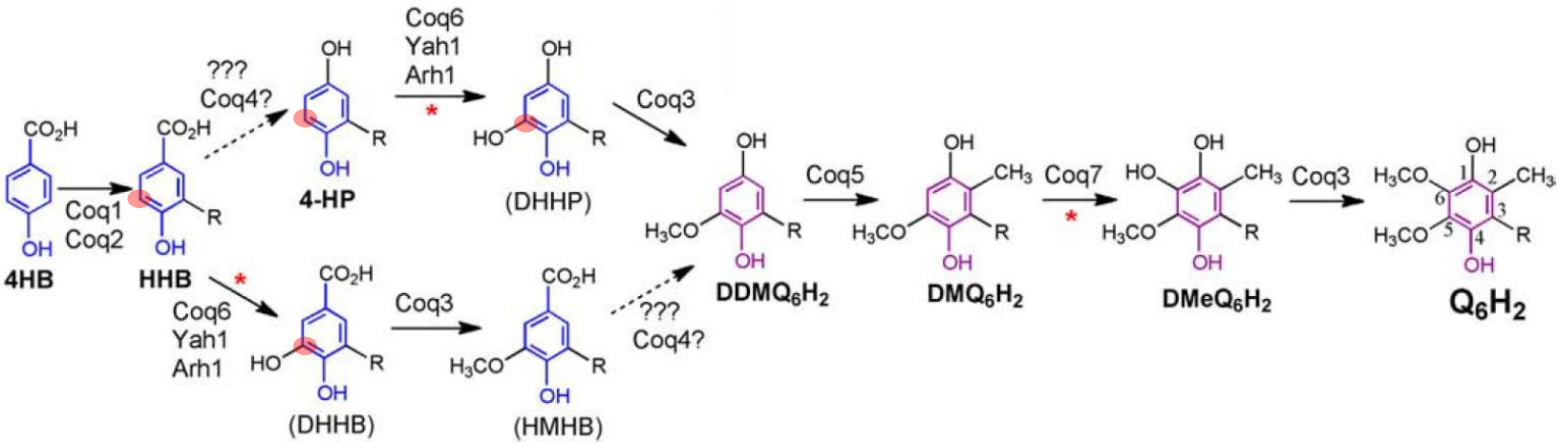

FIG 1.5 *The Q biosynthesis pathway from S. cerevisiae, adapted from "Coenzyme Q supplementation or over-expression of the yeast Coq8 putative kinase stabilizes multi-subunit Coq polypeptide complexes in yeast coq null mutants" (Biochimica and Biophysica Acta 2014).*[18] *The C5 carbon is highlighted in red.*

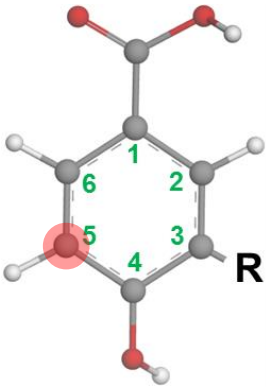

FIG 1.6 *The carbon numbering of the aromatic head of Q biosynthesis intermediates as shown on 4-hydroxybenzoate. R indicates the position occupied by the polyprenyl tail. The C5 carbon is highlighted in red.*

However, the exact attribution of biosynthetic intermediates to each monooxygenase and methyltransferase cannot be completed[22] purely through *in* vivo studies because of the functional (and presumably structural) interdependence of these enzymes in the CoQ synthome. Additionally, at least one enzyme of this pathway, Coq2, displays substrate promiscuity[23], meaning there may not be a single unique order of pathway reactions. This aspect of Q biosynthesis still awaits more complete *in vivo* and *in vitro* characterization of each enzyme and its role in the pathway.

The structural relationship between the Coq proteins is the subject of active research. This work usually combines sub-cellular fractionation, gel purification, affinity purification, and mass spectrometry to determine protein-protein contacts in the protein complex.[24] This protein interaction data is the basis of evolving models of the CoQ synthome, the latest published example[18] of which is presented below in **Figure 1.7**.

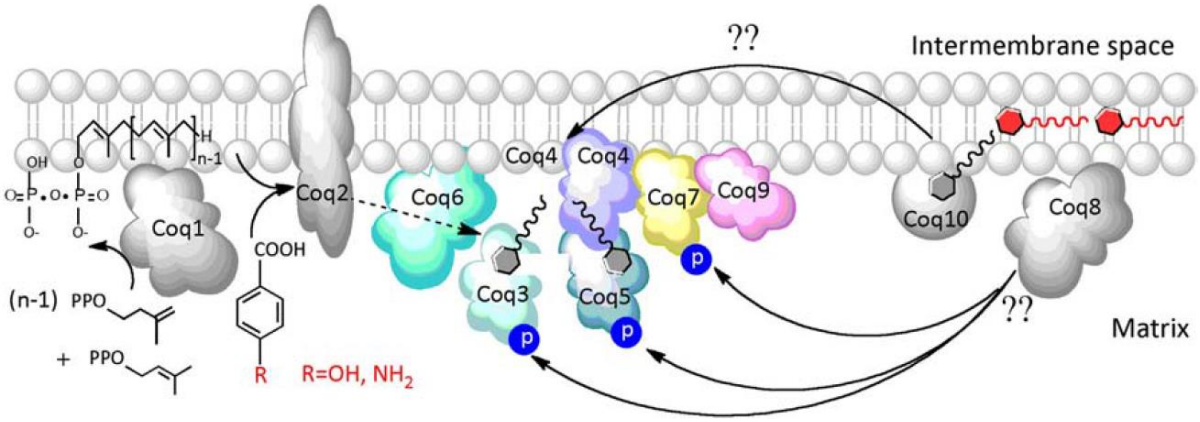

FIG 1.7 *Coq synthome cartoon from "Coenzyme Q supplementation or over-expression of the yeast Coq8 putative kinase stabilizes multi-subunit Coq polypeptides (Biochimica and Biophysica Acta 2014).*[18]

Single gene knockout studies of Coq1-9 typically show accumulation only of the early biosynthesis intermediate 3-hexaprenyl-4-hydroxybenzoic acid (HHB), which requires only the synthesis and attachment of the polyprenyl tail to the aromatic head (in its basal state as derived from chorismate or tyrosine).[25] This indicates that none of the remaining Q biosynthesis proteins are able to function without the presence of the others, implying the requirement of protein-protein interactions to permit individual enzymatic function.

However, the molecular details of this inter-dependency are not clear. Does each enzyme physically require the other proteins for passive structural stability as a pre-condition for enzymatic activity? Are the proteins regulated through other effectors, such as phosphorylation? In such a case, the assembly of the complex would not be strictly necessary for the catalysis on the substrate, but rather for activation of component proteins. Are incomplete Coq synthomes actively degraded by the cell? The answers to these questions have important implications for developing *in vitro* enzymatic activity assays for individual Coq enzymes.

2.2 Individual Coq proteins

In this section we will briefly describe each of the known Q biosynthesis enzymes from *S. cerevisiae* based on the literature. We will also add descriptions of their predicted structures as inferred from close homologs found through a preliminary structural modeling of the Coq proteins using an automated protein modeling server, Phyre2.[26] A more complete description of the preliminary modeling is presented in Annex 1.

Coq1 is a mitochondrial hexaprenyl pyrophosphate synthase. It is a peripheral membrane protein localized to the matrix face of the inner mitochondrial membrane. It is responsible for catalyzing the synthesis of the hexaprenyl tail from isopentenyl pyrophosphate units, one of the first steps of Q biosynthesis. While the structure of Coq1 has not been determined, the structures of several bacterial homologs have been solved, including geranyl diphosphate synthase from *Arabidopsis thaliana* (PDB entry 3AQ0). This enzyme, showing 44% sequence identity to Coq1, and reveals an 8 helix bundle crystallized as a dimer. This is consistent with experimental results indicating that dimerization is necessary to elongate the polyprenyl chain to the target length, which varies by species and is determined by the action of Coq1.[27] The naturally observed length variation of the polyprenyl tail among extant model organisms does not seem to be a critical parameter for the function of Q in the respiratory chain, as complementation of *S. cerevisiae* Coq1 knockouts by orthologs from *S. pombe, R. norvegicus,* and *H. sapiens* restores aerobic respiration.[28] This shows that the isoprenyl tail length is not a determining factor in molecular recognition of Q biosynthesis intermediates by the other Coq proteins.[29]

Coq2 is a mitochondrial polyprenyl transferase responsible for the condensation of the hexaprenyl tail to the aromatic ring (derived from 4-hydroxybenzoate). The structure of *S. cerevisiae* Coq2 itself has not been determined, although the structure of a functional homolog from *Aeropyrum pernix*[30] (21% sequence identity to Coq2) has been solved. This structure, PDB entry 4OD5, reveals a 9 helix bundle with a hydrophobic belt which forms an oligomerization surface in this crystal form. This predicted structural feature makes it likely that this protein is a transmembrane helix bundle[31], consistent with the localization of one of its substrates, the strongly lipophilic polyprenyl chain synthesized by Coq1. As is

likely to be the case for other Q biosynthesis proteins, Coq2 is not strictly sensitive to the length of the polyprenyl chain.[29 32]

Coq3 is a SAM (S-adenosyl methionine, an enzymatic cofactor) dependent O-methyltransferase responsible for the methylation of the C5 hydroxyl group added by Coq6 and the C6 hydroxyl group added by Coq7.[33] The structure of Coq3 has not yet been determined, but the structure of the functional homolog from *Escherichia coli*, UbiG (36% sequence identity to Coq3), has been deposited as PDB entry 4KDC. This structure describes a small globular protein consisting of an eight stranded beta sheet flanked by helix bundles on both outer faces, and is likely to be a soluble protein associated to the matrix face of the IMM, as it co-purifies with Coq4 in digitonin solubilized mitochondrial extracts.

Coq4 is not known to have any enzymatic activity, but its presence is essential for Q biosynthesis. Experiments indicate it co-purifies with Coq3 [34] and is important for maintenance of expression levels of Coq7 and Coq5.[35] The structure of Coq4 has not been solved, although a putative Q biosynthesis homolog from *Nostoc punctiforme* has (PDB entry 3KB4, 19% sequence ID to Coq4). This structure describes a complex bundle of 10 helices which form dimers using a small hydrophobic patch in the crystal. It is likely to be associated with the matrix face of the IMM.

Coq5 is a SAM dependent C-methyltransferase responsible for the methylation of C2. The structure of Coq5 was recently solved[36] with SAM (PDB entry 4OBW) and without SAM (PDB entry 4OBX). It is composed of a seven stranded beta sheet flanked by helices on both outer faces. This enzyme was crystallized in a tetrameric form consisting of a dimer of dimers. Only one of these dimers shows significant contact through a hydrophobic patch, suggesting that the dimer of dimers may be a crystal-induced oligomerization state. Coq5 co-purifies with Coq4, suggesting it is a peripheral membrane protein on the matrix face of the IMM.

Coq6 is an FAD-dependent monooxygenase responsible for the hydroxylation of C5. Since it is the target of the molecular modeling developed in this work, we will describe it in more detail in Section 2.4 “Coq6: Existing experimental data”.

Coq7, also known as Cat5 in many databases and literature sources, is a monooxygenase responsible for the hydroxylation of C6. While the crystal structure of this enzyme has not been resolved, sequence similarity searches consistently find bacterial ferritin-like proteins such as bacterioferritin from *Blastochloris viridis* (20% sequence identity, PDB entry 4AM4). This structure consists of an elongated four helix bundle with a shallow hydrophobic groove used for dimerization and binding a heme molecule. Despite being a very different fold from Coq6, it is likely that Coq7 is reduced by the same multi-protein electron transfer system of Arh1 and Yah1, described below. Coq7 co-purifies with Coq9 and is associated to the matrix face of the IMM.[37]

Coq8 is likely to be a kinase based on sequence analysis which places it in the ABC1 family.[38] Our preliminary bioinformatics analysis with Phyre2 indicate that it has a central kinase domain, flanked on the N-terminal and C-terminal sides by helix-bundle domains which may be involved in protein-protein contacts. The central kinase domain is very similar (43% sequence identity) to that of human mitochondrial ADCK3 whose structure was recently solved[39] as PDB entry 4PED. While the precise

enzymatic role of Coq8 in regulating other Coq proteins through phosphorylation is still unclear, its general presence is essential to Q synthesis. Beyond this, Coq8 overexpression is able to stabilize other Coq proteins in Coq knockout mutants, making it a tool which has allowed the accumulation of previously unobserved Q biosynthesis intermediates.[18]

Coq9 has been shown to interact with Coq7.[40] It has also been implicated in the deaminase activity of Coq6. Its current function is not known. Its fold is similar to TetR DNA binding transcription protein, but it does not appear to have retained nucleic acid binding features, or catalytic activity. It has been co-crystallized with a lipid, suggesting a possible role as a chaperone which desorbs Q biosynthesis intermediates from the membrane for complete processing by the CoQ synthome.[37]

Coq10 encodes a protein with a START-domain,[41] typically used for lipid binding and transport. According to our preliminary bioinformatics analysis it shows 22% sequence identity to the CC1736 protein from *Caulobacter crescentus*, a structure solved as part of a crystallization campaign of the Northeast Structural Genomics Consortium and deposited as PDB entry 1T17. This structure has a fold consisting of a 7 strand twisted beta sheet flanked by two alpha helices on one side. Assays of Q content in Coq10 null mutants shows nearly normal levels of Q, but phenotypes consistent with a non-functional respiratory chain. This suggests that respiration may require protein-facilitated transport of Q after it has been synthesized.

Coq11 is a new addition to the Coq family, first described in 2015.[17] Previously known only as open reading frame YLR290C, it was discovered to be part of the CoQ synthome through tandem affinity purification. While the exact biochemical function is not known, **Coq11 deletion mutants accumulate C1-carboxylated Q biosynthesis intermediates** bearing the polyprenyl chain on C3 and either a hydroxyl group (if cells were supplemented with 4-hydroxybenzoate) or an amino group (if supplemented with 4-aminobenzoate) at the C4 position, as well as drastically reduced quantities of mature ubiquinone. **This implies a role in decarboxylation**. Since failure to decarboxylate in Coq11 knockouts prevents further modification of the aromatic head, it is likely to be an essential step for downstream processing. That is to say, the **other enzymes of the pathway, including Coq6, may not be able to operate on a carboxylated substrate**. Our preliminary bioinformatics analysis finds sequence homology to template structures consisting of a Rossmann-fold, often with a reductase function.

Arh1 and Yah1 have been found to be essential for the *in vivo* activity of Coq6 as the redox system for reducing Coq6.[16] Arh1 and Yah1 are homologous to the better known mammalian mitochondrial electron transfer proteins adrenodoxin reductase (AdxR) and adrenodoxin (Adx), respectively. *Arh1* encodes an adrenodoxin reductase which uses electrons from NAD(P)H to reduce its own flavin adenine dinucleotide (FAD) cofactor. Reduced Arh1 then transfers these electrons to Yah1's iron-sulfur cluster. This reduced Yah1 can then reduce the Coq6 FAD cofactor to yield a reactive Coq6 enzyme. When considered alone, the introduction of the Arh1/Yah1 pair as an electron transfer mechanism for Coq6 may seem more complex than the direct binding of NAD(P)H as a reductant observed in other flavoprotein monooxgenases. However, promoter-controlled disruption of Arh1 and Yah1 activity[16] abrogates both C5 hydroxylation (indicating a loss of function of Coq6) and C6 hydroxylation (indicating a loss of function of Coq7), despite the presence of both the *Coq6* and *Coq7* genes. This suggests that Coq6 and Coq7 depend on the Arh1/Yah1 pair to transfer the reducing equivalents required for ubiquinone biosynthesis. Unlike

the ubiquinone biosynthesis pathway of *E. coli*, which encodes three homologous flavoprotein monooxygenases of the same global fold as Coq6 to perform all of the hydroxylations on the quinone ring, the pathway in *S. cervisiae* uses an iron-sulfur enzyme, Coq7, to perform the C6 hydroxylation. The difference in global fold and redox cofactor between Coq6 and Coq7 likely makes the direct utilization of NAD(P)H as a common reductant for both enzymes impossible. **That is to say, Yah1/Arh1 pair may be the common source of electrons for the essential C5 and C6 hydroxylations of Q biosynthesis**.

2.3 Known structures of Q biosynthesis proteins

The state of structural knowledge of the Coq proteins listed above is summarized in **Table 1.1** below. Since most *Coq* proteins have not been resolved structurally, we present the homologous proteins from bacterial model organisms, typically *E. coli.* This table shows the generally low sequence identity between the *S. cerevisiae* Coq proteins and their orthologs. Low sequence identity (less than 30%, as computed by the Phyre2 server) is generally indicative of a difficult homology modeling scenario.

TABLE 1.1 *A summary of experimentally resolved Q biosynthesis protein structures from S. cerevisiae. The functional homologs from E. coli are shown for comparison. When examples from neither model organism are available we list structurally similar proteins determined through a hidden Markov-model based search. Sequence identities are shown with respect to the S. cerevisiae Coq proteins.*

S. cerevisiae			**E. coli**				**Other homologs**		
Coq protein	**Structure**	**PDB code**	**Ubi protein**	**Structure**	**PDB code**	**Seq. ID**	**PDB code**	**Seq. ID**	**Function**
COQ1	no		ISPB	No			3AQ0	44%	polyprenyl synthase
							3MZV	36%	polyprenyl synthase
							1WY0	35%	polyprenyl synthase
COQ2	no		UBIA	No			4OD5	21%	4-HB octaprenyltransferase
COQ3	no		UBIG	Yes	4KDC	36%			O-methyltransferase
COQ4	no		UBIX	Yes	1SBZ	22.70%			Decarboxylase
			UBID	Yes	2IDB	23.30%			Decarboxylase
COQ5	yes	4OBX	UBIE	No					C2 C-methyltransferase
COQ6	no		UBII	Partial	4K22	22.60%			C5 monooxygenase
COQ7	no		UBIF?	No			4AM4	20%	C6 monooxygenase
COQ8	no		no homolog known				4PED	43%	kinase / synthome anchor
COQ9	no		no homolog known				4RHP	25%	unknown / lipid binding
COQ10	no		no homolog known				1T17	22%	unkown / Q chaperone
COQ11	no		no homolog known				2ZKL	21%	unknown / Rossmann fold
ARH1	no		no homolog known						adrenodoxin reductase
YAH1	yes	2MJD	no homolog known						adrenodoxin (reduces Coq6)

2.4 Coq6: Existing experimental data

2.4.1 Coq6 amino acid sequence

The Coq6 sequence encodes a protein 479 residues long (UniProt accession number P53318). The first 17 residues comprise the mitochondrial signal sequence, giving the mature form of the protein an effective *in vivo* length of 462 residues. While our preliminary bioinformatics survey of the Coq proteins (see Annex 1) indicated that it is of the same global fold as PHBH, Coq6 (and the larger eukaryotic Coq6-family of proteins) differs significantly from known bacterial homologs by the presence of a large (sometimes greater than 50 residue) sequence not found in other enzymes of the same global fold. This structural feature is explored in more depth in Chapter 3 (Construction of homology models and stability screening through molecular dynamics). Despite this, existing sequence studies of Coq6[49] have grouped it with several Class A flavoprotein monooxygenases.[42] The key characteristics of Class A enzymes are that they are encoded by a single gene encoding a single polypeptide chain containing a Rossmann fold domain (used to bind an FAD cofactor), which they reduce with NAD(P)H. Sequence alignment with similar monooxygenases identifies three sequence motifs that are conserved in this class. These three motifs are used for binding the ADP moiety of FAD, binding the ribityl moiety of FAD, and binding NAD(P)H, as shown in **Figure 1.8** below

Coq6 amino acid sequence (479 residues)

```
         10          20          30          40          50
MFFSKVMLTR RILVRGLATA KSSAPKLTDV LIVGGGPAGL TLAASIKNSP
         60          70          80          90         100
QLKDLKTTLV DMVDLKDKLS DFYNSPPDYF TNRIVSVTPR SIHFLENNAG
        110         120         130         140         150
ATLMHDRIQS YDGLYVTDGC SKATLDLARD SMLCMIEIIN IQASLYNRIS
        160         170         180         190         200
QYDSKKDSID IIDNTKVVNI KHSDPNDPLS WPLVTLSNGE VYKTRLLVGA
        210         220         230         240         250
DGFNSPTRRF SQIPSRGWMY NAYGVVASMK LEYPPFKLRG WQRFLPTGPI
        260         270         280         290         300
AHLPMPENNA TLVWSSSERL SRLLLSLPPE SFTALINAAF VLEDADMNYY
        310         320         330         340         350
YRTLEDGSMD TDKLIEDIKF RTEEIYATLK DESDIDEIYP PRVVSIIDKT
        360         370         380         390         400
RARFPLKLTH ADRYCTDRVA LVGDAAHTTH PLAGQGLNMG QTDVHGLVYA
        410         420         430         440         450
LEKAMERGLD IGSSLSLEPF WAERYPSNNV LLGMADKLFK LYHTNFPPVV
        460         470
ALRTFGLNLT NKIGPVKNMI IDTLGGNEK
```

FIG 1.8: *The Coq6 amino acid sequence contains three easily recognizable sequence features: the GxGxxG motif in orange, the GxDGxxx motif in blue, and the GDAxH motif in green. These motifs are involved in FAD binding. The mitochondrial signal sequence is highlighted in beige.*

Coq6 localizes mainly to mitochondria (as revealed by GFP tagging).[43] Subcellular fractionation and membrane destabilization show that mitochondrial Coq6 is found on the matrix side of the inner mitochondrial membrane.[49]

2.4.2 Chemical reactivity: hydroxylation and deamination

Coq6 is an FAD dependent monooxygenase responsible for the hydroxylation of C5 of ubiquinone biosynthesis intermediates, as determined through electrochemical analysis of redox active lipids in yeast cells expressing inactive Coq6 point mutants.[20] Coq6 and its eukaryotic and prokaryotic homologs are primarily known to hydroxylate Q biosynthesis intermediates that are hydroxylated at position C4, such as 4-hydroxyphenol. However, recent experiments of our research group demonstrate that *S. cerevisiae* Coq6 is also capable of processing C4 aminated substrates, effectively deaminating substrates at C4 if necessary and performing a C4 hydroxylation in addition to the nominal C5 hydroxylation.[44] This was principally observed through the *S. cerevisiae* ability to synthesize Q using 4-aminophenol as the source of Q's aromatic head. The deamination activity seems to be related to the last 11 residues of the protein, since truncation of this region abolishes C4 deamination but preserves C5 hydroxylation.[44] When C5 hydroxylation is abolished, either through genetic knockout or inactivating mutation, the resulting non-respiring phenotype can be rescued by the addition of vanillic acid or 3,4-dihydroxbenzoic acid (shown in **Figure 1.9**) both of which furnish an aromatic center already hydroxylated at C5.[20]

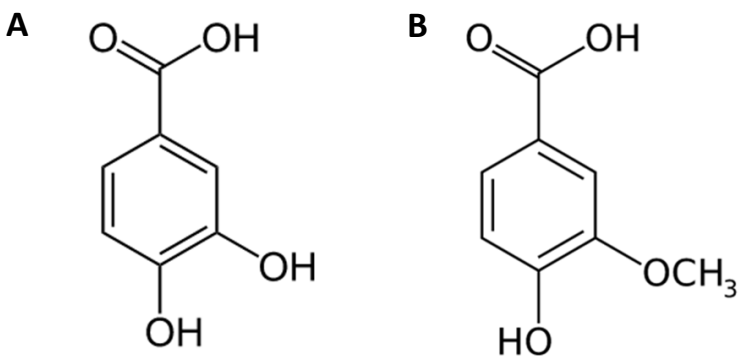

FIG 1.9: *A) 3,4-dihydroxybenzoic acid and B) vanillic acid.*

2.4.3 Protein-protein interactions

Coq6 has been shown to interact with Coq3 and Coq4 through gel filtration of mitochondrial fractions[24] and it has been shown to depend on the presence of Coq9 in order to effect its C4-deaminase activity.[45] Coq6 is also known to co-immunoprecipitate with Coq4, Coq5, Coq7, and Coq9. [40] Together, these results establish that Coq6 is an integral part of the CoQ synthome.

2.4.4 Clinical relevance of Coq6

The Coq6 enzyme is of clinical relevance for humans. Patients with a primary deficiency in Q biosynthesis, a rare recessive disorder, exhibit a variety of heterogenous symptoms including renal and

otological dysfunction[43], mitochondrial dysfunction, encephalomyopathy, ataxia, and cerebellar atrophy.[46] Some cases respond well to oral administration of Q. Several different Coq6 mutations were identified through gene sequencing of clinical cases documenting steroid-resistant nephrotic syndrome with sensorineural deafness. Although the general role of Q in mitochondrial energy production and cell survival are likely to account for some of these deficits, the molecular basis of these dysfunctions are still unclear. In this context, a molecular understanding of Q biosynthesis may allow us to develop new therapies.

One example of this is the study of clinically documented mutant human Coq6 enzymes in the *S. cerevisiae* model system, which was able to demonstrate reduced Q biosynthesis through *in vivo* activity assays.[23] Since Coq6 is the C5 monooxygenase, one aspect of the enzyme's product (if not its precise identity) is already known: it will be hydroxylated at position C5. This suggested that providing Q biosynthesis intermediates already hydroxylated at this position could compensate for the inactivity of the mutant enzymes by providing their product. The molecules chosen as Coq6 product analogs were 3,4-dihydroxybenzoic acid and vanillic acid, which are illustrated in **Figure 1.9**.

Supplementing yeast strains expressing the mutant human *Coq6* genes with these compounds helped improve growth. Molecular knowledge of the Q biosynthesis pathway was essential for this work, as prior work had already established that exogenously supplied aromatic centers with non-standard substitutions (namely an amino group at position C4) can be processed into ubiquinone.[16] Presumably, attachment of the polyprenyl tail is not strictly specific to a single Q biosynthesis intermediate. That is to say, the substrate promiscuity of Coq2 allows cells with a defective Coq6 to assimilate the unprenylated Coq6 products into the biosynthesis pathway.

This molecular understanding suggests that dietary supplementation with vanillic acid could be a treatment for primary Q deficiency and that it is likely to be more effective than supplementation with the pathway's finished product, Q_{10}. The molecular reason for this is well understood: Q_{10} is prenylated and therefore a very hydrophobic molecule, which is a major hindrance to its bio-availability when consumed as a dietary supplement. A biosynthesis intermediate analog without the polyprenyl tail, such as vanillic acid, is much more soluble and therefore much more bioavailable, enabling cells to resume endogenous biosynthesis of their own Q pools.[23]

3. Structures of Q biosynthesis monooxygenases

3.1 Introduction

The Q biosynthesis pathway of yeast contains at least two monooxygenases: Coq6 (responsible for the C5 hydroxylation), Coq7 (responsible for the C6 hydroxylation), and perhaps Coq4 (implicated in the decarboxylation at position C1 as well as possible hydroxylation at this site by our preliminary bioinformatics analysis). In this thesis, we will focus exclusively on Coq6, which builds on the prior experience of our laboratory with the *E. coli* homolog of this enzyme, UbiI (deposited under PDB code 4K22).[47]

The genetic and biochemical studies of our group have shown that the C5 hydroxylation is performed by Coq6, and that this enzyme is likely to be a flavin adenine dinucleotide (FAD) dependent

monooxygenase.[20 48 49] In this thesis, we present a further structure-function characterization of Coq6. Our group has isolated and purified Coq6 for biochemical study (principally through the thesis work of Lucie Gonzalez), but its structure has not been solved experimentally. We thus decided to establish a structure-function characterization on the basis of molecular modeling.

Before proceeding in more detail with the homology-modeled structure of Coq6, the reader will find it helpful to review the state of the art in structure-function characterization of this general class of enzyme, typified by the holotype *para*-hydroxybenzoate hydroxylase (PHBH). This will allow us to introduce molecular structures and chemical functions we are likely to encounter in our modeling of the Coq6 structure.

The similarity between PHBH and Coq6 was revealed in the preliminary molecular modeling survey of Coq proteins described in the previous section and detailed in Annex 1. According to these results, Coq6 shows 15% sequence identity to PHBH, being ranked in 10th position in the Phyre2 template list (shown in Annex 1). While it is not the highest ranking result from the Coq6 template search, it is certainly one of the most valuable because of the wealth of experimental data that has been collected on it through enzymatic characterization and crystallization of the wild-type and many mutants with variations in substrates and products.

3.2 PHBH: Holotype of Class A flavoprotein monooxygenases

A preliminary structural description of Coq6 can be developed through the analysis of a structurally similar enzyme, PHBH, which will introduce the general structure-function features of the global fold (defined as PHBH-like in SCOP, the Structural Classification Of Proteins[50]) likely to be conserved in Coq6. Our sequence-based search for proteins of known structure similar to Coq6 returns PHBH, (with 15% sequence identity to Coq6) which is also one of the most extensively characterized enzymes in biochemistry. The wild-type and many mutants have been studied by detailed *in vitro* activity assays under many variations of pH, substrate type, cofactor type, and cofactor reduction system.[51 52 53 54 55 56 57 58 59 60 87] The global fold of PHBH is interesting for its ability to execute two chemical reactions within a single active site in a single polypeptide chain, using conformational changes to combine an enzymatic cofactor (FAD), an FAD reductant (NADH in the case of PHBH), molecular oxygen, and the substrate in a coordinated sequence to produce a regioselectively hydroxylated product. We will briefly review the general anatomy of the global fold and the catalytic cycle of PHBH focusing on the specific structure-function relationships (including FAD movements) likely to be relevant for our study of Coq6.

3.2.1 PHBH: Global fold and FAD

PHBH is composed of a single polypeptide chain whose global fold produces two distinct structural domains with two distinct functions (see **Figure 1.10 Panels B, C, D**). One domain is mainly composed of a large beta-sheet used to bind the substrate (para-hydroxybenzoate, pHB) as well as to close the active site and exclude bulk solvent, which is an important feature of catalysis. The other domain is mainly composed of a Rossmann-fold and is used to bind FAD in an extended conformation.

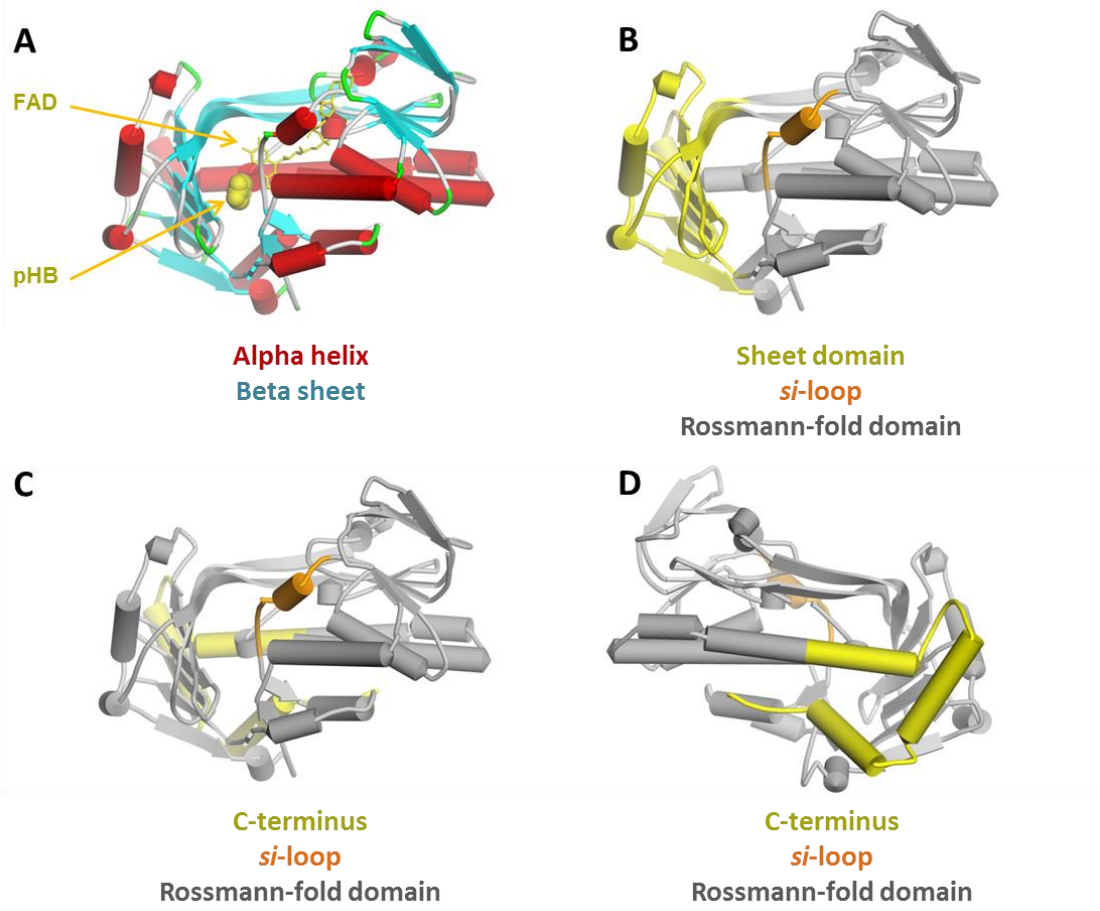

FIG 1.10 *PHBH global anatomy. A) PHBH structure 1PBE colored by secondary structure. The FAD cofactor is in yellow stick. The enzyme's substrate pHB is shown as yellow spheres (omitted for clarity in panels B, C, and D). The enzyme is viewed from the si face, named after the si face of the FAD's isoalloxazine ring. B) PHBH structure colored by functional sub-domains. The Rossmann-fold domain is shown in gray, the beta sheet domain is in yellow, and the si loop is in orange. C) The C-terminus is highlighted in yellow, as seen from the si face, and D) from the re face.*

FAD is the cofactor essential for catalysis in this class of enzymes. The FAD itself is composed of five distinct moieties as illustrated in **Figure 1.11** which have specific interactions with the binding pocket important for molecular motion implicated in catalysis. The adenine, ribose, and pyrophosphate moieties are not directly necessary for catalysis. Rather, they provide a common molecular "handle" to allow the stable binding of the cofactor. The pyrophosphate is followed by a ribityl chain, which terminates with the isoalloxazine ring system. The plane of the isoalloxazine ring is approximately perpendicular to a plane defined by the helices of the FAD binding domain. The isoalloxazine ring plane has two faces, named *si* and *re*, as illustrated in **Figure 1.12B**. The *si* and *re* face nomenclature is derived from the Cahn-Ingold-Prelog priority rules[61] for naming stereocenters. In the case of FAD, the relevant stereocenter for naming the isoalloxazine ring faces is the C2' carbon of the ribityl chain, as indicated in **Figure 1.11**. The

catalysis mediated by FAD occurs at the isoalloxazine ring, which is positioned at the juncture between the FAD binding domain and the substrate binding domain.

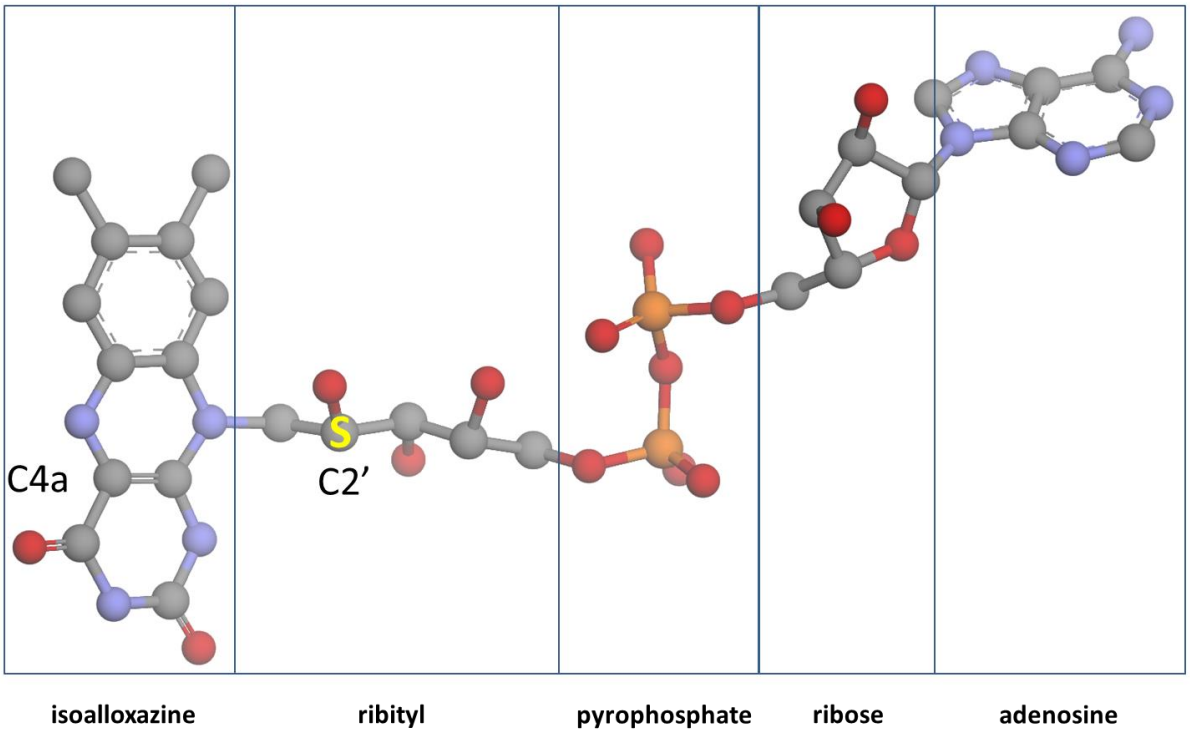

FIG 11: *Anatomy of the FAD cofactor. FAD is composed of five distinct chemical moieties, each providing a specific function for protein binding or catalysis. The asymmetric carbon (labeled C2') nearest the isoalloxazine is used to assign the naming of the* si *and* re *faces of the isoalloxazine ring according to the Cahn-Ingold-Prelog system. Also shown is the C4a carbon which will bear the reactive peroxo group. Hydrogens are omitted for clarity.*

The ribityl chain and the isoalloxazine in particular are loosely bound in the enzyme's pocket, allowing a "vertical" swinging displacement of the isoalloxazine in the plane defined by its ring system. This motion is implied by two crystal structures of PHBH which capture the isoalloxazine in two distinct positions, named *in* (PDB structure 1PBE[62]) and *out* (PDB structure 1PBB[63]), as well as an intermediate conformation (PDB structure 1K0I[54]). The FAD performs catalysis by forming a reactive peroxo-flavin adduct through the addition of molecular oxygen at the C4a carbon. A representative structure of this species is described by PDB structure 2JBV[90], wherein FAD forms a molecular adduct with oxygen upon X-ray illumination of its parent enzyme, choline oxidase. We note that the oxygen is added to the *re* face of the isoalloxazine, and that the formation of this adduct changes the hybridization of the C4a carbon from sp^2 to sp^3, disrupting the planarity of the isoalloxazine and making it much bulkier. These points are resumed in the following **Figure 1.12**.

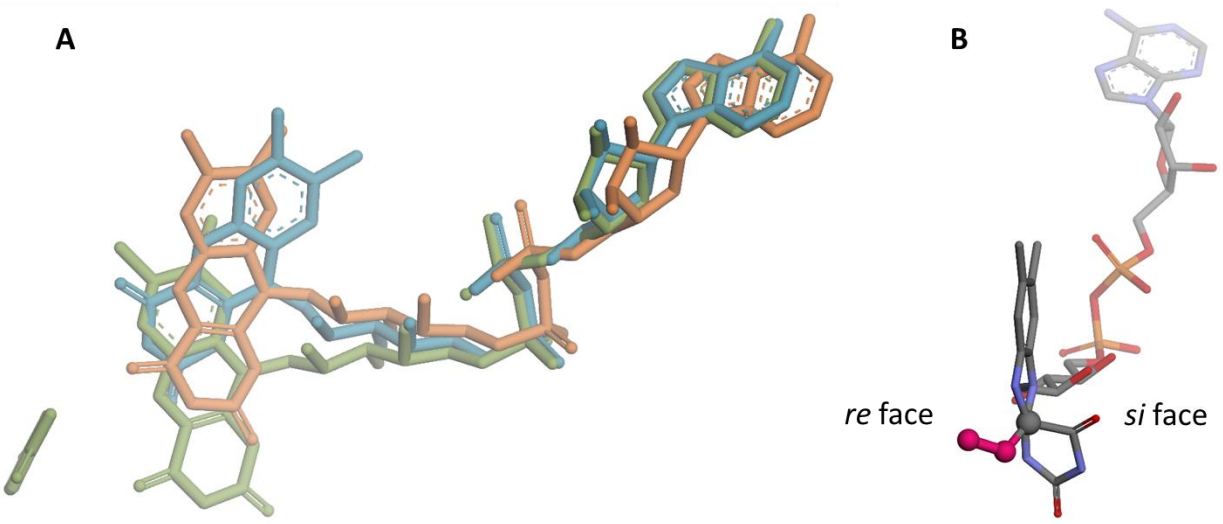

FIG 1.12: A) *Crystallographically determined* in *and* out *conformations of FAD. Green: FAD from PDB structure 1PBE,*[62] *in conformation. The co-crystallized substrate, para-hydroxybenzoate, is shown to the left in stick, the aromatic ring is seen edge-on. Blue: FAD from PDB structure 1PBB,*[63] *out conformation. Orange: FAD from PDB structure 1K0I,*[54] *an intermediate conformation.* B) *2JBV*[90] *structure showing peroxo-flavin adduct forming on the re face of the ring system. The si and re faces of the isoalloxazine are indicated.*

The *si* face of the ring gives its name to an element of secondary structure common to the global fold: the *si* loop shown in orange in **Figure 1.10**. This surface exposed stretch of the polypeptide chain emerges from the first beta-alpha-beta motif of the Rossmann fold. It reaches the surface of the protein and defines part of the *si* face of the FAD binding pocket before plunging back into a buried position in the protein. This stretch of residues is of variable length among the top templates. In 1PBE it is 12 residues long, whereas in Coq6 it is predicted to be 22 residues long. As it is surface exposed and does not pack against any other secondary structure elements, it is quite mobile and often not resolved in crystal structures. When it is resolved, it is often in an alpha helical conformation, as it is in 1PBE. [64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 83]

The *re* face of the ring can be used to designate the other half of the enzyme, extending from the *re* face of the FAD binding pocket to the edge of the protein. This half of the global fold is typically composed of a three strand beta sheet running parallel to the long axis of the FAD. The outermost "edge strand" is initiated just after a short stretch of helix, and is a region implicated in binding NADH in some family members.[54] Similar to the *si* loop, this outermost beta strand is surface exposed, and can be mobile enough to elude resolution by X-ray diffraction in some solved structures.

The PHBH enzyme itself also undergoes motions important for catalysis, transitioning between states known as *open* and *closed*.[57] The 1K0I[54] structure (a single point mutant, R220Q) captures the PHBH structure in its *open* conformation, where the active site is accessible to the bulk solvent, as opposed to structure 1PBE, where it is not. The sequence of conformational transitions for enzyme and cofactor during the catalytic cycle known for PHBH are detailed in the following section. This brief review will

highlight some key features of the catalytic cycle and the structural features associated with these functions. In particular, we will focus on the geometry of the cofactor and substrate in the active site, as it will be an essential reference geometry for analyzing our homology models of Coq6.

To resume, the key features of this global fold are:

1. 2 domains: the Rossmann-fold domain adapted for binding FAD, and the sheet domain adapted for binding substrates.
2. The enzyme structure is centered around the FAD, with the isoalloxazine ring held in a "vertical" plane, and its motion largely confined to sliding of the isoalloxazine within this plane. The structural nomenclature of the isoalloxazine ring system's *si* and *re* faces can be extended to the enzyme as well. This is a convenient structural convention for further studies since the position of the isoalloxazine also defines the active site.
3. Isolation of the active site from bulk solvent during catalysis to prevent the loss of peroxo-flavin's peroxo group as H_2O_2.

These features are described in a series of experimentally solved enzyme-ligand complexes,[54] giving us a series of reference structures for modeling the Coq6 enzyme-ligand system. In the following section we will review the catalytic cycle of PHBH to describe the structures associated with each function.

3.2.2 PHBH: Catalytic cycle

The global fold of PHBH is remarkable in its ability to coordinate the binding of four ligands (FAD, the FAD reductant, molecular oxygen, and the substrate) in a sequential manner in a single polypeptide and within a single active site. The identity of the substrate is verified through a specific substrate deprotonation event. This prevents the wasteful consumption of reducing equivalents on incorrect substrates. The enzyme must also isolate the reactive peroxo-FAD adduct from the bulk solvent to prevent the generation of harmful peroxide species.[82]

In order to provide the molecular complementarity required for each ligand binding event, this single polypeptide chain undergoes several distinct conformational changes over the course of its catalytic cycle, which we will detail here. The catalytic cycle illustrated in **Figure 1.13** can be broken down into 5 steps:

1. Substrate binding
2. Substrate recognition by deprotonation
3. FAD transition to the *out* conformation and FAD reduction by NADH
4. FAD transition to the *in* conformation and formation of a peroxo-flavin adduct
5. Substrate hydroxylation and product release

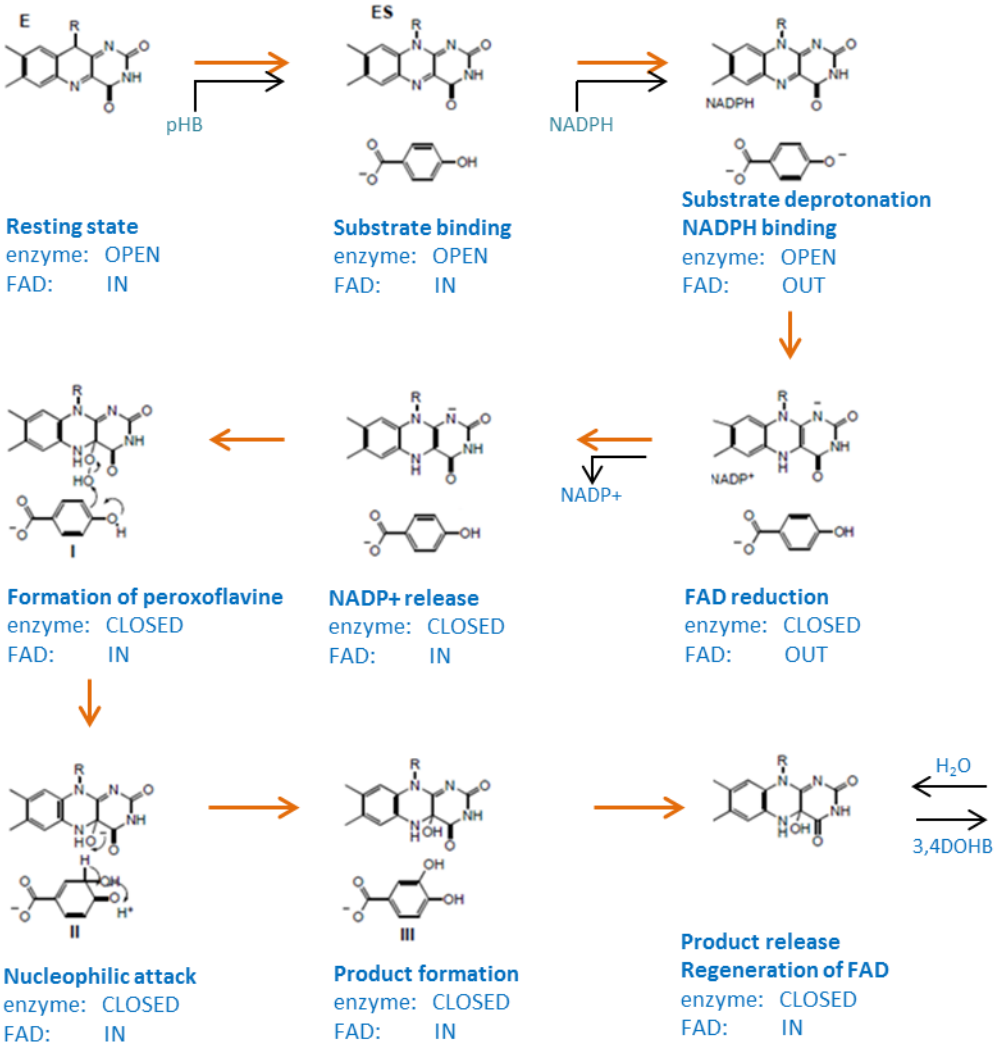

FIG 1.13: Catalytic cycle of PHBH showing the major cofactor and substrate intermediates. Adapted from *Ballou et al (2005).*[60]

1. The cycle begins with the enzyme in its *open* conformation and the FAD in its *in* conformation. In this conformation the active site is open to the bulk solvent, allowing entrance of the substrate (pHB) to the active site. This particular enzyme has an interesting adaptation to cellular energy conservation: it will not consume a reducing equivalent (in the form of NADH) to reduce FAD for catalysis unless it has bound the "correct" substrate – para-hydroxybenzoate.

2. The enzyme confirms substrate identity through deprotonation of para-hydroxybenzoate's phenolic hydrogen. This proton is moved from the substrate to the solvent through a proton transfer network consisting of H72, Y385, Y201, and several crystallographic water molecules.[60] This network of spatially proximal residues enables the enzyme to deprotonate the substrate while isolating it from bulk solvent. Failure to deprotonate the substrate, either through binding an incorrect substrate or mutations

disrupting the proton transfer network, impairs catalysis. At this point the enzyme switches to a *closed* conformation, isolating the active site from the solvent. Deprotonation of the substrate forms a dianionic species, which shifts the FAD to the *out* conformation.

3. The FAD's *out* conformation increases the exposure of the isoalloxazine, opening more accessible volume on the *re* face of the ring system. This accessible volume is likely to be required by the nicotinamide ring of NADH for direct hydride transfer to the isoalloxazine by a direct ring stacking as suggested by the crystal structure of cyclohexanone monooxygenase complexed with FAD and NADPH (PDB entry 3GWD).[83]

4. The reduction of FAD is followed by its transition back to the *in* conformation, where it can form the peroxo-flavin adduct, ready to hydroxylate the substrate.

5. The substrate is hydroxylated, and the enzyme switches to the *open* conformation for product release.

While the general sequence of events in the PHBH catalytic cycle are likely to be recapitulated in Coq6, some details may well differ. A more particular feature of PHBH, the proton-transfer network essential to substrate deprotonation, was not found to be reproduced in our Coq6 model. However, given that the function of the deprotonation step is to verify the identity of the substrate, we infer that it is probably not essential for Coq6 to employ this mechanism. This is because Coq6 is an obligate member of a specialized protein complex apposed to the inner mitochondrial membrane and is likely to face a much more limited set of possible substrates, making such discrimination unnecessary.

3.3 Monooxygenases: Existing computational studies

The wealth of experimentally solved structures of wild-type and mutant PHBH (and other flavoprotein monooxygenases) in complex with many substrate and cofactor variants have produced a large base of experimental atomic coordinates for computational studies. These indicate that an atomic resolution structure of the Coq6 enzyme can help us understand substrate binding and catalysis phenomena at a similar level of detail. Here we will briefly review five examples from the literature of the interplay between modeling and experiment in the characterization and rational design of FPMOs. First, we will see an example of the rational re-design of ligand binding in phenylacetone monooxygenase based on homology models of the enzyme.[85] In the second example we will see molecular dynamics applied to the investigation of the proton transfer network in PHBH.[88] The third example describes accessible volume calculations performed on the PHBH structure.[89] The fourth example describes the combined use of homology modeling, molecular dynamics, and docking to test computed substrate affinity to experimental dissociation constants.[91] Finally, we include an example of quantum mechanical modeling applied to the PHBH system to show the utility of atomic resolution protein structures.[93 94]

These examples serve to highlight three types of calculations important for the study of Coq6: computational redesign of ligand binding, molecular dynamics, accessible volumes, and substrate docking. These examples establish a precedent for the modeling strategy and techniques we will apply to Coq6.

3.3.1 Computational redesign of ligand binding based on homology models

PHBH is a type of flavoprotein monooxygenase (FPMO), an important class of enzymes in industrial chemistry, as they allow the stereoselective monooxygenation of substrates. These wild-type enzymes are excellent starting points for modification towards creating enzymes for processing industrial substrates through directed evolution or structure based rational design. Structural knowledge of these enzymes has been used to alter substrate specificity and product stereochemistry through rational designed mutations.[84]

An example of this is the rational redesign of the thermostable phenylacetone monooxygenase (PAMO) by Pazmino *et al*.[85] PAMO's thermostability makes it a good candidate for an industrial biocatalysis enzyme, but it accepts only a small number of mainly aromatic substrates: phenylacetone, benzylacetone, alpha-methylphenylacetone, 4-hydroxyacetophenone, 2-dodecanone, bicyclohept-2-en-6-one, and methyl-4-tolylsulfide.[86] In order to expand the substrate scope treatable by this enzyme, particularly towards aliphatics, the authors turned to a homologous enzyme with greater substrate scope, cyclopentanone monooxygenase (CPMO), but lower stability. The authors identified key residues in the PAMO active site which were *not* conserved in CPMO, reasoning that these residues in CPMO are the molecular basis for accepting more diverse substrates. These positions in PAMO were mutated to their CPMO counterparts in various combinations and tested for activity, revealing a single point mutation which allowed the binding of a novel substrate.

A key feature of the PAMO work is that the identification of substrate binding residues was done through the comparison of an experimentally solved PAMO structure (determined by crystallography) and a computationally predicted CPMO structure (created by homology modeling). This is an example of the practical utility of an experimentally validated homology model in identifying substrate binding residues and designing mutations. A similar but more detailed strategy for modeling Coq6 enzyme-substrate interactions is developed in Chapter 2 (Computational strategy and methods).

3.3.2 Molecular dynamics studies

Inspection of early PHBH crystal structures identified a network of titratable residues and crystallographic water molecules connecting the active site to the protein surface and bulk solvent: H72, Y385, and Y201.[87] The purpose of this network is to transfer a proton away from the substrate to the solvent while preventing direct contact of the solvent to the active site. This was proposed to be accomplished by proton hopping between the titratable residues in the network. This was corroborated by the reduced reactivities of substrates which cannot be deprotonated and of mutations disrupting the proton transfer network. The direction of the proton flow depends primarily on the orientation of the side-chain of H72. While crystal structures can give us the coordinates of the proton transfer network, they cannot tell us about the dynamic behavior of the residues involved, which is particularly important for determining the rotameric state (and therefore orientation) of H72. Molecular dynamics is a method uniquely capable of exploring these conformational states and transitions at atomic resolution. In the case of PHBH, standard molecular dynamics simulations of PHBH in different titration states enabled investigators[88] to sample conformations accessible from the crystal structure. In the case of this proton transfer network the functionally relevant protein movements are governed primarily by sidechain rotations. Therefore, standard molecular dynamics simulations on relatively short timescales (sub-

microsecond) enabled the authors to sample many relevant conformations for the PHBH system and provide structural explanations for the differing enzyme reactivities caused by different substrates and different mutations to the enzyme.

This is a good example of MD used to simulate functional behavior in this class of enzymes. Similar calculations will be used in the structural analysis and conformational sampling of Coq6 molecular models, as described in Chapter 3 (Construction of Coq6 homology models and stability screening through molecular dynamics).

3.3.3 Accessible volume calculation

Molecular modeling of PHBH began shortly after the resolution of the 1PHH crystal structure with the *Analysis of the active site of the flavoprotein p-Hydroxybenzoate hydroxylase and some ideas with respect to its reaction mechanism* by Schreuder *et al* (1990).[89] In this work, the authors use molecular modeling to explore the possible positions for the distal oxygen of the flavin-dioxygen adduct through rotation of the O-C4a bond. They found three sterically favorable positions, one of which could correspond to a catalytic positioning, and another which could be compatible with reduction by NADPH. The third position demonstrated an accessible volume on the *re* face side of isoalloxazine ring. NADPH is likely to appose its hydride bearing nicotinamide ring to the *re* face of the isoalloxazine. The accessible volume also makes it likely that the dioxygen adduct forms on the *re* face of the enzyme, since it involves conversion of the planar sp^2 hybridized C4a carbon to a tetrahedral sp^3 hybridized form. This tetrahedral geometry is bulkier than the planar aromatic geometry of the FAD in its resting state, requiring more accessible volume which can only be found on the *re* face.

This computationally developed hypothesis for the PHBH peroxo-flavin geometry was crystallographically confirmed with the resolution of the choline oxidase structure 2JBV.[90] While not a Class A flavoenzyme like PHBH, choline oxidase also uses a flavin cofactor to perform its reaction. The 2JBV[90] crystallization construct formed an oxygenated adduct on the *re* side of the flavin C4a atom under X-ray illumination, providing a first structure of a peroxo-flavin species co-crystallized in a protein.

This first peroxo-flavin modeling work on PHBH structure 1PHH dates from 1990 and is a prototypical example of the importance of accessible volume calculations. More modern accessible volume calculations will be used in characterizing the Coq6 active site as described in Chapter 2 (Computational Strategy and Methods) and as applied in Chapter 3 (Developing the hypothesis of a substrate access channel).

3.3.4 Substrate docking

Studying enzyme-substrate interactions through molecular dynamics and substrate docking in FPMOs also has precedent in the 2006 study of Feenstra *et al* [91] on styrene monooxygenase and a series of possible substrates. The authors first created a homology model of styrene monooxygenase (since it had not been experimentally solved at the time). They then used molecular dynamics to refine their model and perform conformational sampling. Conformations derived from dynamics were then used for substrate docking and binding affinity calculation.

This work is an interesting example of the linking of three modeling techniques (homology modeling, molecular dynamics, and substrate docking) in a concerted strategy for generating structure-function hypotheses for an enzyme in the absence of any crystallographic data. In the current work, we elaborate a similar strategy as described in Chapter 2 (Computational strategy and methods). This includes the application of substrate docking into the Coq6 homology model as described in Chapter 5 (Selection of Coq6 models through molecular dynamics and substrate docking).

3.3.5 QM/MM modeling

Several enzyme-cofactor-substrate complexes have been crystallized for PHBH. However, such complexes are not guaranteed to describe the coordinates of an enzymatically competent system. A good example of this is the PDB entry 1K0J,[54] in which PHBH is co-crystallized with FAD and its reductant NADH. However, the NADH is bound with its adenine ring apposed to the upper edge of the isoalloxazine *si* face. This conformation is unlikely to be compatible with FAD reduction since the hydride of the reduced NADH is borne on the nicotinamide ring – a position 18 Å away from the isoalloxazine in the 1K0J conformation.

Crystals of the PHBH enzyme-substrate complex (without NADH, as in PDB entry 1PBE) and the enzyme-product complex (as in PDB entry 1PHH[92]) are a better representation of a chemically reactive complex, making these coordinates a plausible starting point for simulations of the chemical reaction itself. This can be studied computationally with methods that combine molecular dynamics (to simulate larger scale protein motions) and quantum mechanics (to simulate bond breaking and formation during catalysis) in the PHBH system.[93] [94] While we will not develop QM/MM simulations for Coq6 in this work, these simulations on the PHBH system are important because they tell us that geometric features of enzyme-substrate interactions captured in the PHBH crystal structures are competent for catalysis. We will use these geometric features as guides in the docking-based approach of modeling enzyme-substrate interactions for the Coq6 system, as described in **Chapter 4** (Selection of Coq6 models through molecular dynamics and substrate docking). The most relevant measurement of molecular geometry in the PHBH system is the distance between the FAD isoalloxazine C4a carbon and the target carbon on the substrate to be hydroxylated: 4.38Å.

4. Discussion

4.1 Challenges of studying the Coq system and the value added of molecular modeling

While continuing work in the field is likely to provide better enzymatic and structural characterization of the pathway enzymes, there are four fundamental challenges that may limit the isolated study of individual Q biosynthesis enzymes. These are: i) enzyme solubility, ii) substrate solubility, iii) the enzymes' redox system, and iv) the functional interdependence of the Coq proteins in the CoQ synthome.

Together, these challenges suggest that creating *in vitro* constructs for Coq protein crystallization (and activity assays) will be more difficult than for the case of cytosolic proteins that do not form obligate multi-protein complexes. That is to say, the *in vivo* study of the Coq enzymes, including Coq6, may

progress much faster than their experimental coordinates can be acquired. In this context of challenging structural characterization, molecular modeling has significant value added in developing residue- or atomic-resolution structure-function hypotheses. In this section we will briefly describe the four fundamental challenges of the Coq system and how molecular modeling can contribute to addressing them.

4.1.1 Enzyme solubility

The isolation and purification of enzymes for *in vitro* characterization requires a good control of enzyme solubility, particularly at the high concentrations used for crystallization. However, because Coq proteins form part of a larger protein complex in direct contact with the inner mitochondrial membrane, their surfaces have likely evolved to form specific protein-protein and protein-membrane contacts. That is to say, it is likely that many of the Coq proteins have evolved to *not* be soluble in aqueous solution, a property which has hindered the structural resolution of both Coq6 and its bacterial homolog UbiI. This makes molecular modeling of the enzymes relevant to improving the experimental enzyme purification process as well as providing predictive molecular models. Molecular modeling of protein surface properties, such as electrostatic and aromatic surfaces, can be used to rationalize and modify the solution behavior observed for these proteins. This contribution is described in greater detail in Chapter 6 (Research perspectives).

4.1.2 Substrate solubility

In vitro characterization of catalysis requires a substrate, and poor aqueous solubility is a property shared by all Q biosynthesis intermediates. This is because attachment of the polyprenyl tail is one of the first steps in the pathway. Therefore, even if it is possible to create a soluble enzyme construct, providing it with an appropriate substrate in an aqueous *in vitro* assay may prove difficult. This makes molecular modeling of enzyme-substrate interactions a valuable technique which can generate structure-function hypotheses before structural resolution of the enzyme-substrate complexes. Molecular modeling of these interactions can also inform the design of substrate analogs which can strike a balance between solubility and reactivity. This contribution is described in greater detail in Chapter 4 (Selection of models by substrate docking).

Indeed, the poor aqueous solubility of Q biosynthesis intermediates may have been a contributing factor to the evolution of the CoQ synthome as a membrane-associated protein complex, since the attachment of the polyprenyl tail at the beginning of the pathway is likely to make desorption from the membrane energetically unfavorable for all Q biosynthesis intermediates.

4.1.3 Enzyme redox systems

The Coq enzymes require reducing equivalents to perform catalysis on their substrates. While the Q biosynthesis enzymes themselves have been identified, their redox systems are not fully understood. A good example of this is the discovery of the requirement of the Arh1 and Yah1 proteins for Coq6 function.[16] Despite having the sequence motifs and predicted structure of a Class A flavoprotein (whose members generally obtain their reducing equivalents from direct binding of NADH or NADPH), Coq6

cannot reduce its FAD cofactor this way, but must get its electrons from the Arh1/Yah1 electron transfer system.

When enzyme reduction is accomplished through binding of small molecules (such as NADPH), molecular docking is a technique which can describe the enzyme-ligand interaction. When enzyme is reduction is accomplished through a protein-protein interaction we must turn to protein-protein docking. This technique is more difficult to successfully apply to systems without experimentally determined coordinates for the protein partners. This describes the case of Coq6, which seems to require Arh1 and Yah1 for reduction. The protein-protein binding aspect of Coq6 is not treated in the present work.

4.1.4 Enzyme interdependence

Finally, the interdependence of the Coq proteins for stability, activity, and substrate access makes *in vitro* studies of the pathway more difficult. However, molecular modeling can be to applied design experiments and extract more conclusive *in vivo* results. Knowledge of the active site residues of each Coq enzyme, accessible through homology modeling, can allow us to formulate an experimental strategy for determining substrate-enzyme attribution through the systematic design and testing of catalytically inactive Coq mutants. This combination can bring detailed structural knowledge and modification of Coq proteins to *in vivo* studies, where the natural system has already addressed the problems of protein solubility, substrate solubility, and enzyme reduction. This is described further in the Chapter 6 (Discussions and Perspectives).

5. Conclusion

Questions addressed by the present work

A long-term goal of the laboratory (the Laboratory of Chemistry of Biological Processes, or LCPB) is to study and characterize a number of proteins of the Q biosynthesis system and possibly determine the structure of any larger-scale protein complexes they may form. The laboratory's study of this pathway more specifically concerns Coq6, the C5 monooxygenase and seeks to characterize this protein with *in vivo* studies (conducted at the University of Grenoble), *in vitro* studies (conducted at the Collège de France), and structural *in silico* studies, (also conducted at the Collège de France).

In this context, the goal of this thesis is to study Coq6 computationally and establish a structure-function characterization of this enzyme-cofactor-substrate system. The structural characterization seeks to answer three main questions:

1. What is the atomic-resolution structure of Coq6?
2. How does Coq6 bind its cofactor?
3. How does Coq6 bind its substrate?

In this work we develop and analyze atomic resolution molecular models of Coq6 and its ligands. These molecular models are used to develop structure-function hypotheses for Coq6, as well as the rational design of mutations to test them. We will describe this development in a stepwise fashion.

Chapter 2 will describe the computational strategy and methods we will use from a theoretical perspective. We will first translate the three main questions of the present study into general molecular modeling tasks. These tasks are then further translated into specific techniques of molecular modeling. We will then compose these techniques into a larger strategy for modeling Coq6. Given the low sequence identity of Coq6 to the possible templates, the general goal of the strategy is to generate and test several possible models of Coq6 before simulating their interaction with a model substrate.

Chapter 3 will describe the application of the strategy developed in Chapter 2. We will first describe the construction of a set of Coq6 homology models using three independent methods. We will then test their structural stability using molecular dynamics.

Chapter 4 will describe the selection of Coq6 models based on attempts at docking a model substrate into the Coq6 active site. First, we will analyze our homology models to identify the active site and any possible substrate binding sites to derive geometric descriptors with which we will analyze the molecular dynamics trajectories created in Chapter 3.

Chapter 5 will describe the *in silico* testing of the substrate access channel in a yeast Coq6 mutant modeled from human clinical literature. We will further test the putative substrate access channel by using our homology model to rationally design additional mutations to block the channel. Finally, these predictions will be tested by *in vivo* activity assays.

Finally, in Chapter 6 we will present the conclusions of the current work and perspectives on the application of molecular modeling to the study of Coq6 and the entire Q biosynthesis pathway.

References

[1] Bentinger, Magnus, Michael Tekle, and Gustav Dallner. "Coenzyme Q – Biosynthesis and Functions." *Biochemical and Biophysical Research Communications* 396, no. 1 (May 21, 2010): 74–79. doi:10.1016/j.bbrc.2010.02.147.

[2] Ernster, Lars, and Gustav Dallner. "Biochemical, Physiological and Medical Aspects of Ubiquinone Function." *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, Nobel symposium 90: Mitochondrial diseases, 1271, no. 1 (May 24, 1995): 195–204. doi:10.1016/0925-4439(95)00028-3.

[3] Collins, M D, and D Jones. "Distribution of Isoprenoid Quinone Structural Types in Bacteria and Their Taxonomic Implication." *Microbiological Reviews* 45, no. 2 (June 1981): 316–54.

[4] Nowicka, Batrycze, and Jerzy Kruk. "Occurrence, Biosynthesis and Function of Isoprenoid Quinones." *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1797, no. 9 (September 2010): 1587–1605. doi:10.1016/j.bbabio.2010.06.007.

[5] Turunen, Mikael, Jerker Olsson, and Gustav Dallner. "Metabolism and Function of Coenzyme Q." *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1660, no. 1–2 (January 28, 2004): 171–99. doi:10.1016/j.bbamem.2003.11.012.

[6] Kanehisa, Minoru, and Susumu Goto. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28, no. 1 (January 1, 2000): 27–30.

[7] Nohl, H., and L. Gille. "The Existence and Significance of Redox-Cycling Ubiqunone in Lysosomes." Protoplasma, 2001.

[8] Maroz, Andrej, Robert F. Anderson, Robin A. J. Smith, and Michael P. Murphy. "Reactivity of Ubiquinone and Ubiquinol with Superoxide and the Hydroperoxyl Radical: Implications for in Vivo Antioxidant Activity." *Free Radical Biology and Medicine* 46, no. 1 (January 1, 2009): 105–9. doi:10.1016/j.freeradbiomed.2008.09.033.

[9] Clarke, Catherine F., Amy C. Rowat, and James W. Gober. "Osmotic Stress: Is CoQ a Membrane Stabilizer?" *Nature Chemical Biology* 10, no. 4 (April 2014): 242–43. doi:10.1038/nchembio.1478.

[10] Haines, Thomas H. "Do Sterols Reduce Proton and Sodium Leaks through Lipid Bilayers?" *Progress in Lipid Research* 40, no. 4 (July 2001): 299–324. doi:10.1016/S0163-7827(01)00009-1.

[11] Quinn, Peter J. "Lipid–lipid Interactions in Bilayer Membranes: Married Couples and Casual Liaisons." *Progress in Lipid Research* 51, no. 3 (July 2012): 179–98. doi:10.1016/j.plipres.2012.01.001.

[12] Sévin, Daniel C., and Uwe Sauer. "Ubiquinone Accumulation Improves Osmotic-Stress Tolerance in Escherichia Coli." *Nature Chemical Biology* 10, no. 4 (April 2014): 266–72. doi:10.1038/nchembio.1437.

[13] Hauß, Thomas, Silvia Dante, Thomas H. Haines, and Norbert A. Dencher. "Localization of Coenzyme Q10 in the Center of a Deuterated Lipid Membrane by Neutron Diffraction." *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1710, no. 1 (November 15, 2005): 57–62. doi:10.1016/j.bbabio.2005.08.007.

[14] Tzagoloff, A., & Dieckmann, C. L. (1990). PET genes of Saccharomyces cerevisiae. *Microbiological Reviews*, *54*(3), 211–225.

[15] Cui, Tie-Zhong, and Makoto Kawamukai. "Coq10, a Mitochondrial Coenzyme Q Binding Protein, Is Required for Proper Respiration in Schizosaccharomyces Pombe." *FEBS Journal* 276, no. 3 (February 1, 2009): 748–59. doi:10.1111/j.1742-4658.2008.06821.x.

[16] Pierrel, Fabien, Olivier Hamelin, Thierry Douki, Sylvie Kieffer-Jaquinod, Ulrich Mühlenhoff, Mohammad Ozeir, Roland Lill, and Marc Fontecave. "Involvement of Mitochondrial Ferredoxin and Para-Aminobenzoic Acid in Yeast Coenzyme Q Biosynthesis." *Chemistry & Biology* 17, no. 5 (May 28, 2010): 449–59. doi:10.1016/j.chembiol.2010.03.014.

[17] Allan, Christopher M., Agape M. Awad, Jarrett S. Johnson, Dyna I. Shirasaki, Charles Wang, Crysten E. Blaby-Haas, Sabeeha S. Merchant, Joseph A. Loo, and Catherine F. Clarke. "Identification of Coq11, a New Coenzyme Q Biosynthetic Protein in the CoQ-Synthome in Saccharomyces Cerevisiae." *Journal of Biological Chemistry*, January 28, 2015, jbc.M114.633131. doi:10.1074/jbc.M114.633131.

[18] He, Cuiwen H., Letian X. Xie, Christopher M. Allan, UyenPhuong C. Tran, and Catherine F. Clarke. "Coenzyme Q Supplementation or over-Expression of the Yeast Coq8 Putative Kinase Stabilizes Multi-Subunit Coq Polypeptide Complexes in Yeast Coq Null Mutants." *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1841, no. 4 (April 2014): 630–44. doi:10.1016/j.bbalip.2013.12.017.

[19] Meganathan, R. "Ubiquinone Biosynthesis in Microorganisms." *FEMS Microbiology Letters* 203, no. 2 (September 1, 2001): 131–39. doi:10.1111/j.1574-6968.2001.tb10831.x.

[20] Ozeir, Mohammad, Ulrich Mühlenhoff, Holger Webert, Roland Lill, Marc Fontecave, and Fabien Pierrel. "Coenzyme Q Biosynthesis: Coq6 Is Required for the C5-Hydroxylation Reaction and Substrate Analogs Rescue Coq6 Deficiency." *Chemistry & Biology* 18, no. 9 (September 23, 2011): 1134–42. doi:10.1016/j.chembiol.2011.07.008.

[21] Meurant, Gerard. *Vitamins and Hormones: Advances in Research and Applications Volume 40*. Academic Press, 1983.

[22] Tran, UyenPhuong C., and Catherine F. Clarke. "Endogenous Synthesis of Coenzyme Q in Eukaryotes." *Mitochondrion* 7, Supplement (June 2007): S62–71. doi:10.1016/j.mito.2007.03.007.

[23] Doimo, Mara, Eva Trevisson, Rannar Airik, Marc Bergdoll, Carlos Santos-Ocaña, Friedhelm Hildebrandt, Placido Navas, Fabien Pierrel, and Leonardo Salviati. "Effect of Vanillic Acid on COQ6 Mutants Identified in Patients with Coenzyme Q10 Deficiency." *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842, no. 1 (January 2014): 1–6. doi:10.1016/j.bbadis.2013.10.007.

[24] Marbois, Beth, Peter Gin, Kym F. Faull, Wayne W. Poon, Peter T. Lee, Jeff Strahan, Jennifer N. Shepherd, and Catherine F. Clarke. "Coq3 and Coq4 Define a Polypeptide Complex in Yeast Mitochondria for the Biosynthesis of Coenzyme Q." *Journal of Biological Chemistry* 280, no. 21 (May 27, 2005): 20231–38. doi:10.1074/jbc.M501315200.

[25] Xie, Letian X., Mohammad Ozeir, Jeniffer Y. Tang, Jia Y. Chen, Sylvie-Kieffer Jaquinod, Marc Fontecave, Catherine F. Clarke, and Fabien Pierrel. "Overexpression of the Coq8 Kinase in Saccharomyces Cerevisiae Coq Null Mutants Allows for Accumulation of Diagnostic Intermediates of the Coenzyme Q6 Biosynthetic Pathway." *Journal of Biological Chemistry* 287, no. 28 (July 6, 2012): 23571–81. doi:10.1074/jbc.M112.360354.

[26] Riccardo, Riccardo M. Bennett Lovsey, Alex D. Herbert, Lawrence A. Kelley, and Michael J. E. Sternberg. "Exploring the Extremes of Sequence/structure Space with Ensemble Fold Recognition in the Program Phyre." *Proteins* 70, no. 3 (February 15, 2008): 611–25. doi:10.1002/prot.21688.

[27] Kainou, Tomohiro, Kazunori Okada, Kengo Suzuki, Tsuyoshi Nakagawa, Hideyuki Matsuda, and Makoto Kawamukai. "Dimer Formation of Octaprenyl-Diphosphate Synthase (IspB) Is Essential for Chain Length Determination of Ubiquinone." *Journal of Biological Chemistry* 276, no. 11 (March 16, 2001): 7876–83. doi:10.1074/jbc.M007472200.

[28] Gin, Peter, and Catherine F. Clarke. "Genetic Evidence for a Multi-Subunit Complex in Coenzyme Q Biosynthesis in Yeast and the Role of the Coq1 Hexaprenyl Diphosphate Synthase." *Journal of Biological Chemistry* 280, no. 4 (January 28, 2005): 2676–81. doi:10.1074/jbc.M411527200.

[29] Okada, Kazunori, Tomohiro Kainou, Hideyuki Matsuda, and Makoto Kawamukai. "Biological Significance of the Side Chain Length of Ubiquinone in Saccharomyces Cerevisiae." *FEBS Letters* 431, no. 2 (July 17, 1998): 241–44. doi:10.1016/S0014-5793(98)00753-4.

[30] Cheng, Wei, and Weikai Li. "Structural Insights into Ubiquinone Biosynthesis in Membranes." *Science* 343, no. 6173 (February 21, 2014): 878–81. doi:10.1126/science.1246774.

[31] Quinzii, Catarina, Ali Naini, Leonardo Salviati, Eva Trevisson, Plácido Navas, Salvatore DiMauro, and Michio Hirano. "A Mutation in Para-Hydroxybenzoate-Polyprenyl Transferase (COQ2) Causes Primary Coenzyme Q10 Deficiency." *The American Journal of Human Genetics* 78, no. 2 (February 2006): 345–49. doi:10.1086/500092.

[32] Ashby, M. N., S. Y. Kutsunai, S. Ackerman, A. Tzagoloff, and P. A. Edwards. "COQ2 Is a Candidate for the Structural Gene Encoding Para-Hydroxybenzoate:polyprenyltransferase." *Journal of Biological Chemistry* 267, no. 6 (February 25, 1992): 4128–36.

[33] Poon, Wayne W., Robert J. Barkovich, Adam Y. Hsu, Adam Frankel, Peter T. Lee, Jennifer N. Shepherd, David C. Myles, and Catherine F. Clarke. "Yeast and Rat Coq3 and Escherichia Coli UbiG Polypeptides Catalyze

Both O-Methyltransferase Steps in Coenzyme Q Biosynthesis." *Journal of Biological Chemistry* 274, no. 31 (July 30, 1999): 21665–72. doi:10.1074/jbc.274.31.21665.

[34] Marbois, Beth, Peter Gin, Melissa Gulmezian, and Catherine F. Clarke. "The Yeast Coq4 Polypeptide Organizes a Mitochondrial Protein Complex Essential for Coenzyme Q Biosynthesis." *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1791, no. 1 (January 2009): 69–75. doi:10.1016/j.bbalip.2008.10.006.

[35] Belogrudov, Grigory I., Peter T. Lee, Tanya Jonassen, Adam Y. Hsu, Peter Gin, and Catherine F. Clarke. "Yeast COQ4 Encodes a Mitochondrial Protein Required for Coenzyme Q Synthesis." *Archives of Biochemistry and Biophysics* 392, no. 1 (August 1, 2001): 48–58. doi:10.1006/abbi.2001.2448.

[36] Dai, Ya-Nan, Kang Zhou, Dong-Dong Cao, Yong-Liang Jiang, Fei Meng, Chang-Biao Chi, Yan-Min Ren, Yuxing Chen, and Cong-Zhao Zhou. "Crystal Structures and Catalytic Mechanism of the C-Methyltransferase Coq5 Provide Insights into a Key Step of the Yeast Coenzyme Q Synthesis Pathway." *Acta Crystallographica. Section D, Biological Crystallography* 70, no. Pt 8 (August 2014): 2085–92. doi:10.1107/S1399004714011559.

[37] Lohman, Danielle C., Farhad Forouhar, Emily T. Beebe, Matthew S. Stefely, Catherine E. Minogue, Arne Ulbrich, Jonathan A. Stefely, et al. "Mitochondrial COQ9 Is a Lipid-Binding Protein That Associates with COQ7 to Enable Coenzyme Q Biosynthesis." *Proceedings of the National Academy of Sciences of the United States of America* 111, no. 44 (November 4, 2014): E4697–4705. doi:10.1073/pnas.1413128111.

[38] Leonard, C. J., L. Aravind, and E. V. Koonin. "Novel Families of Putative Protein Kinases in Bacteria and Archaea: Evolution of the 'Eukaryotic' Protein Kinase Superfamily." *Genome Research* 8, no. 10 (October 1998): 1038–47.

[39] Stefely, Jonathan A., Andrew G. Reidenbach, Arne Ulbrich, Krishnadev Oruganty, Brendan J. Floyd, Adam Jochem, Jaclyn M. Saunders, et al. "Mitochondrial ADCK3 Employs an Atypical Protein Kinase-like Fold to Enable Coenzyme Q Biosynthesis." *Molecular Cell* 57, no. 1 (January 8, 2015): 83–94. doi:10.1016/j.molcel.2014.11.002.

[40] Hsieh, Edward J., Peter Gin, Melissa Gulmezian, UyenPhuong C. Tran, Ryoichi Saiki, Beth N. Marbois, and Catherine F. Clarke. "Saccharomyces Cerevisiae Coq9 Polypeptide Is a Subunit of the Mitochondrial Coenzyme Q Biosynthetic Complex." *Archives of Biochemistry and Biophysics* 463, no. 1 (July 1, 2007): 19–26. doi:10.1016/j.abb.2007.02.016.

[41] Barros, Mario H., Alisha Johnson, Peter Gin, Beth N. Marbois, Catherine F. Clarke, and Alexander Tzagoloff. "The Saccharomyces Cerevisiae COQ10 Gene Encodes a START Domain Protein Required for Function of Coenzyme Q in Respiration." *Journal of Biological Chemistry* 280, no. 52 (December 30, 2005): 42627–35. doi:10.1074/jbc.M510768200.

[42] van Berkel, W. J. H., N. M. Kamerbeek, and M. W. Fraaije. "Flavoprotein Monooxygenases, a Diverse Class of Oxidative Biocatalysts." *Journal of Biotechnology* 124, no. 4 (August 5, 2006): 670–89. doi:10.1016/j.jbiotec.2006.03.044.

[43] Heeringa, Saskia F., Gil Chernin, Moumita Chaki, Weibin Zhou, Alexis J. Sloan, Ziming Ji, Letian X. Xie, et al. "COQ6 Mutations in Human Patients Produce Nephrotic Syndrome with Sensorineural Deafness." *The Journal of Clinical Investigation* 121, no. 5 (May 2011): 2013–24. doi:10.1172/JCI45693.

[44] Ozeir, Mohammad, Ludovic Pelosi, Alexandre Ismail, Caroline Mellot-Draznieks, Marc Fontecave, and Fabien Pierrel. "Coq6 Is Responsible for the C4-Deamination Reaction in Coenzyme Q Biosynthesis in Saccharomyces Cerevisiae." *The Journal of Biological Chemistry*, August 10, 2015. doi:10.1074/jbc.M115.675744.

[45] He, Cuiwen H., Dylan S. Black, Theresa P. T. Nguyen, Charles Wang, Chandra Srinivasan, and Catherine F. Clarke. "Yeast Coq9 Controls Deamination of Coenzyme Q Intermediates That Derive from Para-Aminobenzoic Acid." *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1851, no. 9 (September 2015): 1227–39. doi:10.1016/j.bbalip.2015.05.003.

[46] Rötig, Agnés, Eeva-Liisa Appelkvist, Vanna Geromel, Dominique Chretien, Noman Kadhom, Patrick Edery, Marc Lebideau, et al. "Quinone-Responsive Multiple Respiratory-Chain Dysfunction due to Widespread Coenzyme Q10 Deficiency." *The Lancet* 356, no. 9227 (July 29, 2000): 391–95. doi:10.1016/S0140-6736(00)02531-9.

[47] Chehade, Mahmoud Hajj, Laurent Loiseau, Murielle Lombard, Ludovic Pecqueur, Alexandre Ismail, Myriam Smadja, Béatrice Golinelli-Pimpaneau, et al. "ubiI, a New Gene in Escherichia Coli Coenzyme Q Biosynthesis, Is Involved in Aerobic C5-Hydroxylation." *Journal of Biological Chemistry* 288, no. 27 (July 5, 2013): 20085–92. doi:10.1074/jbc.M113.480368.

[48] Clarke, Catherine F. "Coq6 Hydroxylase: Unmasked and Bypassed." *Chemistry & Biology* 18, no. 9 (September 23, 2011): 1069–70. doi:10.1016/j.chembiol.2011.09.006.

[49] Gin, Peter, Adam Y. Hsu, Steven C. Rothman, Tanya Jonassen, Peter T. Lee, Alexander Tzagoloff, and Catherine F. Clarke. "The Saccharomyces Cerevisiae COQ6 Gene Encodes a Mitochondrial Flavin-Dependent Monooxygenase Required for Coenzyme Q Biosynthesis." *Journal of Biological Chemistry* 278, no. 28 (July 11, 2003): 25308–16. doi:10.1074/jbc.M303234200.

[50] Fox, Naomi K., Steven E. Brenner, and John-Marc Chandonia. "SCOPe: Structural Classification of Proteins—extended, Integrating SCOP and ASTRAL Data and Classification of New Structures." *Nucleic Acids Research* 42, no. D1 (January 1, 2014): D304–9. doi:10.1093/nar/gkt1240.

[51] Gatti, D. L., B. A. Palfey, M. S. Lah, B. Entsch, V. Massey, D. P. Ballou, and M. L. Ludwig. "The Mobile Flavin of 4-OH Benzoate Hydroxylase." *Science (New York, N.Y.)* 266, no. 5182 (October 7, 1994): 110–14.

[52] Lah, M. S., B. A. Palfey, H. A. Schreuder, and M. L. Ludwig. "Crystal Structures of Mutant Pseudomonas Aeruginosa P-Hydroxybenzoate Hydroxylases: The Tyr201Phe, Tyr385Phe, and Asn300Asp Variants." *Biochemistry* 33, no. 6 (February 15, 1994): 1555–64.

[53] Ortiz-Maldonado, M., D. Gatti, D. P. Ballou, and V. Massey. "Structure-Function Correlations of the Reaction of Reduced Nicotinamide Analogues with P-Hydroxybenzoate Hydroxylase Substituted with a Series of 8-Substituted Flavins." *Biochemistry* 38, no. 50 (December 14, 1999): 16636–47.

[54] Wang, Jian, Mariliz Ortiz-Maldonado, Barrie Entsch, Vincent Massey, David Ballou, and Domenico L. Gatti. "Protein and Ligand Dynamics in 4-Hydroxybenzoate Hydroxylase." *Proceedings of the National Academy of Sciences* 99, no. 2 (January 22, 2002): 608–13. doi:10.1073/pnas.022640199.

[55] Ortiz-Maldonado, M., S. M. Aeschliman, D. P. Ballou, and V. Massey. "Synergistic Interactions of Multiple Mutations on Catalysis during the Hydroxylation Reaction of P-Hydroxybenzoate Hydroxylase: Studies of the Lys297Met, Asn300Asp, and Tyr385Phe Mutants Reconstituted with 8-Cl-Flavin." *Biochemistry* 40, no. 30 (July 31, 2001): 8705–16.

[56] Ortiz-Maldonado, Mariliz, Lindsay J. Cole, Sara M. Dumas, Barrie Entsch, and David P. Ballou. "Increased Positive Electrostatic Potential in P-Hydroxybenzoate Hydroxylase Accelerates Hydroxylation but Slows Turnover." *Biochemistry* 43, no. 6 (February 17, 2004): 1569–79. doi:10.1021/bi030193d.

[57] Cole, Lindsay J., Barrie Entsch, Mariliz Ortiz-Maldonado, and David P. Ballou. "Properties of P-Hydroxybenzoate Hydroxylase When Stabilized in Its Open Conformation." *Biochemistry* 44, no. 45 (November 15, 2005): 14807–17. doi:10.1021/bi0512142.

[58] Ortiz-Maldonado, Mariliz, Barrie Entsch, and David P. Ballou. "Oxygen Reactions in P-Hydroxybenzoate Hydroxylase Utilize the H-Bond Network during Catalysis." *Biochemistry* 43, no. 48 (December 7, 2004): 15246–57. doi:10.1021/bi048115t.

[59] Entsch, Barrie, Lindsay J. Cole, and David P. Ballou. "Protein Dynamics and Electrostatics in the Function of P-Hydroxybenzoate Hydroxylase." *Archives of Biochemistry and Biophysics* 433, no. 1 (January 1, 2005): 297–311. doi:10.1016/j.abb.2004.09.029.

[60] Ballou, David P., Barrie Entsch, and Lindsay J. Cole. "Dynamics Involved in Catalysis by Single-Component and Two-Component Flavin-Dependent Aromatic Hydroxylases." *Biochemical and Biophysical Research Communications* 338, no. 1 (December 9, 2005): 590–98. doi:10.1016/j.bbrc.2005.09.081.

[61] Prelog, Vladlmir, and Günter Helmchen. "Basic Principles of the CIP-System and Proposals for a Revision." *Angewandte Chemie International Edition in English* 21, no. 8 (August 1, 1982): 567–83. doi:10.1002/anie.198205671.

[62] Schreuder, Herman A., Peter A. J. Prick, Rik K. Wierenga, Gerrit Vriend, Keith S. Wilson, Wim G. J. Hol, and Jan Drenth. "Crystal Structure of the p-Hydroxybenzoate Hydroxylase-Substrate Complex Refined at 1.9

Å Resolution: Analysis of the Enzyme-Substrate and Enzyme-Product Complexes." *Journal of Molecular Biology* 208, no. 4 (August 20, 1989): 679–96. doi:10.1016/0022-2836(89)90158-7.

[63] Schreuder, H. A., A. Mattevi, G. Obmolova, K. H. Kalk, W. G. Hol, F. J. van der Bolt, and W. J. van Berkel. "Crystal Structures of Wild-Type P-Hydroxybenzoate Hydroxylase Complexed with 4-aminobenzoate,2,4-Dihydroxybenzoate, and 2-Hydroxy-4-Aminobenzoate and of the Tyr222Ala Mutant Complexed with 2-Hydroxy-4-Aminobenzoate. Evidence for a Proton Channel and a New Binding Mode of the Flavin Ring." *Biochemistry* 33, no. 33 (August 23, 1994): 10161–70.

[64] Ukaegbu, Uchechi E., Auric Kantz, Michelle Beaton, George T. Gassner, and Amy C. Rosenzweig. "Structure and Ligand Binding Properties of the Epoxidase Component of Styrene Monooxygenase." *Biochemistry* 49, no. 8 (March 2, 2010): 1678–88. doi:10.1021/bi901693u.

[65] Leys, D., A. S. Tsapin, K. H. Nealson, T. E. Meyer, M. A. Cusanovich, and J. J. Van Beeumen. "Structure and Mechanism of the Flavocytochrome c Fumarate Reductase of Shewanella Putrefaciens MR-1." *Nature Structural Biology* 6, no. 12 (December 1999): 1113–17. doi:10.1038/70051.

[66] Iverson, Tina M., César Luna-Chavez, Laura R. Croal, Gary Cecchini, and Douglas C. Rees. "Crystallographic Studies of the Escherichia Coli Quinol-Fumarate Reductase with Inhibitors Bound to the Quinol-Binding Site." *The Journal of Biological Chemistry* 277, no. 18 (May 3, 2002): 16124–30. doi:10.1074/jbc.M200815200.

[67] Yankovskaya, Victoria, Rob Horsefield, Susanna Törnroth, César Luna-Chavez, Hideto Miyoshi, Christophe Léger, Bernadette Byrne, Gary Cecchini, and So Iwata. "Architecture of Succinate Dehydrogenase and Reactive Oxygen Species Generation." *Science (New York, N.Y.)* 299, no. 5607 (January 31, 2003): 700–704. doi:10.1126/science.1079605.

[68] Bamford, V., P. S. Dobbin, D. J. Richardson, and A. M. Hemmings. "Open Conformation of a Flavocytochrome c3 Fumarate Reductase." *Nature Structural Biology* 6, no. 12 (December 1999): 1104–7. doi:10.1038/70039.

[69] Malito, Enrico, Andrea Alfieri, Marco W. Fraaije, and Andrea Mattevi. "Crystal Structure of a Baeyer-Villiger Monooxygenase." *Proceedings of the National Academy of Sciences of the United States of America* 101, no. 36 (September 7, 2004): 13157–62. doi:10.1073/pnas.0404538101.

[70] Tsuge, Hideaki, Ryushi Kawakami, Haruhiko Sakuraba, Hideo Ago, Masashi Miyano, Kenji Aki, Nobuhiko Katunuma, and Toshihisa Ohshima. "Crystal Structure of a Novel FAD-, FMN-, and ATP-Containing L-Proline Dehydrogenase Complex from Pyrococcus Horikoshii." *The Journal of Biological Chemistry* 280, no. 35 (September 2, 2005): 31045–49. doi:10.1074/jbc.C500234200.

[71] Madej, M. Gregor, Hamid R. Nasiri, Nicole S. Hilgendorff, Harald Schwalbe, and C. Roy D. Lancaster. "Evidence for Transmembrane Proton Transfer in a Dihaem-Containing Membrane Protein Complex." *The EMBO Journal* 25, no. 20 (October 18, 2006): 4963–70. doi:10.1038/sj.emboj.7601361.

[72] Obiero, Josiah, Vanessa Pittet, Sara A. Bonderoff, and David A. R. Sanders. "Thioredoxin System from Deinococcus Radiodurans." *Journal of Bacteriology* 192, no. 2 (January 2010): 494–501. doi:10.1128/JB.01046-09.

[73] Yeh, Joanne I., Unmesh Chinte, and Shoucheng Du. "Structure of Glycerol-3-Phosphate Dehydrogenase, an Essential Monotopic Membrane Enzyme Involved in Respiration and Metabolism." *Proceedings of the National Academy of Sciences of the United States of America* 105, no. 9 (March 4, 2008): 3280–85. doi:10.1073/pnas.0712331105.

[74] Osawa, Takuo, Koichi Ito, Hideko Inanaga, Osamu Nureki, Kozo Tomita, and Tomoyuki Numata. "Conserved Cysteine Residues of GidA Are Essential for Biogenesis of 5-Carboxymethylaminomethyluridine at tRNA Anticodon." *Structure (London, England: 1993)* 17, no. 5 (May 13, 2009): 713–24. doi:10.1016/j.str.2009.03.013.

[75] Muraki, Norifumi, Daisuke Seo, Tomoo Shiba, Takeshi Sakurai, and Genji Kurisu. "Asymmetric Dimeric Structure of Ferredoxin-NAD(P)+ Oxidoreductase from the Green Sulfur Bacterium Chlorobaculum Tepidum: Implications for Binding Ferredoxin and NADP+." *Journal of Molecular Biology* 401, no. 3 (August 20, 2010): 403–14. doi:10.1016/j.jmb.2010.06.024.

[76] Sasaki, Daisuke, Masahiro Fujihashi, Yuki Iwata, Motomichi Murakami, Tohru Yoshimura, Hisashi Hemmi, and Kunio Miki. "Structure and Mutation Analysis of Archaeal Geranylgeranyl Reductase." *Journal of Molecular Biology* 409, no. 4 (June 17, 2011): 543–57. doi:10.1016/j.jmb.2011.04.002.

[77] Ruggiero, Alessia, Mariorosario Masullo, Maria Rosaria Ruocco, Pasquale Grimaldi, Maria Angela Lanzotti, Paolo Arcari, Adriana Zagari, and Luigi Vitagliano. "Structure and Stability of a Thioredoxin Reductase from Sulfolobus Solfataricus: A Thermostable Protein with Two Functions." *Biochimica Et Biophysica Acta* 1794, no. 3 (March 2009): 554–62. doi:10.1016/j.bbapap.2008.11.011.

[78] Ukaegbu, Uchechi E., Auric Kantz, Michelle Beaton, George T. Gassner, and Amy C. Rosenzweig. "Structure and Ligand Binding Properties of the Epoxidase Component of Styrene Monooxygenase." *Biochemistry* 49, no. 8 (March 2, 2010): 1678–88. doi:10.1021/bi901693u.

[79] Fu, Guoxing, Hongling Yuan, Congran Li, Chung-Dar Lu, Giovanni Gadda, and Irene T. Weber. "Conformational Changes and Substrate Recognition in Pseudomonas Aeruginosa D-Arginine Dehydrogenase." *Biochemistry* 49, no. 39 (October 5, 2010): 8535–45. doi:10.1021/bi1005865.

[80] Franceschini, Stefano, Hugo L. van Beek, Alessandra Pennetta, Christian Martinoli, Marco W. Fraaije, and Andrea Mattevi. "Exploring the Structural Basis of Substrate Preferences in Baeyer-Villiger Monooxygenases: Insight from Steroid Monooxygenase." *The Journal of Biological Chemistry* 287, no. 27 (June 29, 2012): 22626–34. doi:10.1074/jbc.M112.372177.

[81] S. Montersino, D. Tischler. “Catalytic and Structural Features of Flavoprotein Hydroxylases and Epoxidases.” *Advanced Synthesis & Catalysis* 353 (2011): 2301–19.

[82] Brender, Jeffrey R., Joe Dertouzos, David P. Ballou, Vincent Massey, Bruce A. Palfey, Barrie Entsch, Duncan G. Steel, and Ari Gafni. “Conformational Dynamics of the Isoalloxazine in Substrate-Free P-Hydroxybenzoate Hydroxylase: Single-Molecule Studies.” *Journal of the American Chemical Society* 127, no. 51 (December 28, 2005): 18171–78. doi:10.1021/ja055171o.

[83] Mirza, I. Ahmad, Brahm J. Yachnin, Shaozhao Wang, Stephan Grosse, Hélène Bergeron, Akihiro Imura, Hiroaki Iwaki, Yoshie Hasegawa, Peter C. K. Lau, and Albert M. Berghuis. “Crystal Structures of Cyclohexanone Monooxygenase Reveal Complex Domain Movements and a Sliding Cofactor.” *Journal of the American Chemical Society* 131, no. 25 (July 1, 2009): 8848–54. doi:10.1021/ja9010578.

[84] Zhang, Zhi-Gang, Loreto P. Parra, and Manfred T. Reetz. “Protein Engineering of Stereoselective Baeyer–Villiger Monooxygenases.” *Chemistry – A European Journal* 18, no. 33 (August 13, 2012): 10160–72. doi:10.1002/chem.201202163.

[85] Pazmiño, Daniel E. Torres, Radka Snajdrova, Daniela V. Rial, Marko D. Mihovilovic, and Marco W. Fraaije. “Altering the Substrate Specificity and Enantioselectivity of Phenylacetone Monooxygenase by Structure-Inspired Enzyme Redesign.” *Advanced Synthesis & Catalysis* 349, no. 8–9 (June 4, 2007): 1361–68. doi:10.1002/adsc.200700045.

[86] Fraaije, Marco W., Jin Wu, Dominic P. H. M. Heuts, Erik W. van Hellemond, Jeffrey H. Lutje Spelberg, and Dick B. Janssen. “Discovery of a Thermostable Baeyer-Villiger Monooxygenase by Genome Mining.” *Applied Microbiology and Biotechnology* 66, no. 4 (January 2005): 393–400. doi:10.1007/s00253-004-1749-5.

[87] Frederick, K. K., D. P. Ballou, and B. A. Palfey. “Protein Dynamics Control Proton Transfers to the Substrate on the His72Asn Mutant of P-Hydroxybenzoate Hydroxylase.” *Biochemistry* 40, no. 13 (April 3, 2001): 3891–99.

[88] Gatti, D. L., B. Entsch, D. P. Ballou, and M. L. Ludwig. “pH-Dependent Structural Changes in the Active Site of P-Hydroxybenzoate Hydroxylase Point to the Importance of Proton and Water Movements during Catalysis.” *Biochemistry* 35, no. 2 (January 16, 1996): 567–78. doi:10.1021/bi951344i.

[89] Schreuder, H. A., W. G. Hol, and J. Drenth. “Analysis of the Active Site of the Flavoprotein P-Hydroxybenzoate Hydroxylase and Some Ideas with Respect to Its Reaction Mechanism.” *Biochemistry* 29, no. 12 (March 27, 1990): 3101–8.

[90] Quaye, Osbourne, George T. Lountos, Fan, Allen M. Orville, and Giovanni Gadda. “Role of Glu312 in Binding and Positioning of the Substrate for the Hydride Transfer Reaction in Choline Oxidase†,‡.” *Biochemistry* 47, no. 1 (January 1, 2008): 243–56. doi:10.1021/bi7017943.

[91] Feenstra, K. Anton, Karin Hofstetter, Rolien Bosch, Andreas Schmid, Jan N. M. Commandeur, and Nico P. E. Vermeulen. "Enantioselective Substrate Binding in a Monooxygenase Protein Model by Molecular Dynamics and Docking." *Biophysical Journal* 91, no. 9 (November 1, 2006): 3206–16. doi:10.1529/biophysj.106.088633.

[92] Schreuder, H. A., J. M. van der Laan, W. G. Hol, and J. Drenth. "Crystal Structure of P-Hydroxybenzoate Hydroxylase Complexed with Its Reaction Product 3,4-Dihydroxybenzoate." *Journal of Molecular Biology* 199, no. 4 (February 20, 1988): 637–48.

[93] Senn, Hans Martin, Stephan Thiel, and Walter Thiel. "Enzymatic Hydroxylation in P-Hydroxybenzoate Hydroxylase: A Case Study for QM/MM Molecular Dynamics." *Journal of Chemical Theory and Computation* 1, no. 3 (May 1, 2005): 494–505. doi:10.1021/ct049844p.

[94] Ridder, Lars, Adrian J Mulholland, Ivonne M. C. M Rietjens, and Jacques Vervoort. "Combined Quantum Mechanical and Molecular Mechanical Reaction Pathway Calculation for Aromatic Hydroxylation by P-Hydroxybenzoate-3-Hydroxylase." *Journal of Molecular Graphics and Modelling* 17, no. 3–4 (June 1999): 163–75. doi:10.1016/S1093-3263(99)00027-3.

Chapter 2 Computational strategy and methods

1. Introduction

The goal of this thesis on the homology modeling of Coq6 is to establish a structure-function characterization of this enzyme-cofactor-substrate system. This structure-function characterization seeks to answer three main questions defined at the end of Chapter 1 and restated here:

1. What is the atomic-resolution structure of Coq6?
2. How does Coq6 bind its cofactor?
3. How does Coq6 bind its substrate?

These are fundamentally questions of molecular structure, which are best answered in terms of atomic coordinates of the enzyme, its cofactor, and substrate. These coordinates typically come from molecular structures solved through X-ray diffraction or NMR. However, the laboratory's prior experience with the *E. coli* homolog of Coq6, UbiI, indicated that this particular enzyme-ligand system might prove difficult to purify and crystallize for X-ray diffraction. Therefore, we decided to investigate the structure of Coq6 with molecular modeling in parallel with experimental efforts in order to have an alternative source of atomic coordinates from which to generate structure-function hypotheses.

We will use several methods of molecular modeling to produce and analyze a molecular model of Coq6. In this chapter we will describe the materials (computing resources) and methods (theories and software implementations) applied to answering these questions. The individual methods will be presented from their theoretical bases, followed by brief notes on their application to Coq6. For several modeling tasks, the relative difficulty of Coq6 as a modeling target will require the use of some alternative methods not normally necessary for easier modeling targets. The challenges related to the homology modeling of Coq6 required the formulation of a larger modeling strategy. We will first describe the strategy in its general form. We will then present the component methods: homology modeling, molecular dynamics, accessible volume calculations, and docking calculations.

2. Strategy and methods

2.1 From questions to techniques

Answering the three main questions of this study typically requires atomic-resolution structures of the Coq6-FAD-substrate molecular complexes. In the absence of an experimental structure of the enzyme, we decided to use molecular modeling to answer these questions. However, because the atomic coordinates in this study will be generated with molecular modeling, there is the possibility of significant uncertainty in the coordinates. Therefore we add a fourth question: can the predicted enzyme-substrate interactions be verified experimentally? This sequence of questions can be answered through the execution of specific tasks. The mapping between questions and tasks is presented in **Table 2.1** below.

TABLE 2.1: *Fundamental questions of the thesis and the tasks necessary to answer them.*

Question	Task	Technique
1) What is the structure of Coq6?	Create enzyme models	Homology modeling
	Analyze model stability	Molecular dynamics
2) How does Coq6 bind its cofactor?	ID binding site and place cofactor	Template structure study
		Accessible volume calculation
3) How does Coq6 bind its substrate?	ID binding site and place substrate	Residue conservation analysis
		Accessible volume calculation
		Substrate docking
4) Can this be tested experimentally?	Rational design of mutants to modify substrate interactions	Homology modeling
		Molecular dynamics
		Accessible volume calculation
		Substrate docking
	Experimental assay of wild-type and mutant activity	

This sequence of tasks forms the basis of the molecular modeling strategy we will apply in this thesis for the computational study of Coq6. These techniques are listed in their required order of execution (with the exception of evolutionary residue conservation analysis, which can be performed immediately after homology modeling, (or even before, operating on just the Coq6 sequence). We will briefly resume these techniques before describing them in more detail:

Homology modeling is the process of constructing a molecular model of a protein of unknown structure on the basis of a similar protein of known structure.

Molecular dynamics is a physical chemistry simulation technique which can describe the shapes and movements of atoms and molecules through the numerical resolution of Newton's equations of motion. A particular feature of the molecules of our system is that they are able to assume many different shapes, called conformations. Molecules are so small and light that their structures are continuously in motion, assuming many different conformations. This means that even a single isolated molecule is better represented as a collection of conformations rather than as a single static object. Molecular dynamics is a method which allows us to sample collections (or ensembles) of conformations available to the molecules of interest. This will be particularly important for substrate docking, where we will try to find complementary shapes for the Coq6 enzyme and its substrate in order to investigate details of their interactions.

Substrate docking is a technique for simulating how two molecules can fit together, such as enzyme and substrate. In our case, because the enzyme is a molecule much larger than the substrate, we will need to first define a region of the enzyme where the substrate is likely to bind. This substrate binding region is unknown for the Coq6 system, and so we will combine the results of two techniques to identify this region: **evolutionary residue conservation** and **accessible volume calculation**. Evolutionary residue conservation is a technique based on multiple-sequence alignments which reveal which residues are conserved and therefore likely to be functionally important. Accessible volume calculations are a type of

geometry analysis technique which can find regions in a protein structure where a substrate could have the room to bind.

2.2 From techniques to strategy

The specific case of Coq6 presents additional challenges which will inform the composition of these basic techniques into a larger strategy. These challenges are summarized below:

- The low sequence identity between Coq6 and its closest structural templates (<30%)
- Coq6 has large regions with no experimentally solved templates
- The absence of experimentally solved homologous enzyme-substrate complexes

The first point can be addressed through the use of special template search methods designed to perform well at finding distant homologs. Knowledge of specific functional homolog structures (also ubiquinone biosynthesis hydroxylases) combined with knowledge of some fundamental principles of protein structure enabled us to make specific knowledge-based rational choices in template selection that can be more functionally relevant to our specific target, Coq6. Low sequence identity between target and template also makes sequence alignment more difficult, which can be overcome by using a multiple sequence alignment methods.

The second point can be addressed through the creation of multiple homology models using different methods and coordinate sources to generate possible structures for the unknown regions. We can discriminate among these models by testing their structural stability through MD simulations. These simulations calculate the potential energy of each system (meaning the enzyme, its cofactor, and surrounding solvent) and produce a trajectory of each system through its phase space. These trajectories can also be used to perform more detailed analyses of the conformations sampled during the simulation. This is essential for addressing the third challenge, which is the lack of knowledge of how Coq6 and its substrate fit together. We will address this through molecular docking of the substrate, a ubiquinone biosynthesis intermediate, into our multiple models of the Coq6 enzyme. Substrate docking aims to predict the bound conformation of the substrate in the active site of the enzyme. This requires a structural definition of the Coq6 substrate binding site, which is not completely known *a priori* from any homologous enzyme-substrate complexes.

We will then identify the substrate binding site by cross-referencing a geometric analysis of the Coq6 model with evolutionary sequence conservation. Once we have identified a possible substrate binding site, we will use molecular docking to simulate the enzyme-substrate interaction. Docking, which is the matching of complementary molecular shapes (in our case, enzyme and substrate), is complicated by the inherent flexibility of these molecular systems.

Interactions between molecules (in our case, enzyme and ligands) therefore depend not just on their composition, but also on their detailed conformations. This makes it necessary to explore the different conformations available to each partners. In order to account for molecular motion, we will use a technique called molecular dynamics to explore the conformations available to the enzyme, while the substrate docking algorithm includes methods for exploring the much more limited conformational

space of the substrate. The combination of these methods aims to generate complementary molecular shapes which will fit together.

In order to address these three challenges, we will further compose the techniques listed into a larger modeling strategy consisting of four parts: A) building homology models (through manual and automated methods (I-TASSER and ROBETTA); B) analyzing homology models through MD (to identify possible substrate binding regions); C) substrate docking (to functionally test these regions); D) *in silico* rational design and *in vivo* testing of mutants to test these substrate access regions. This is resumed in more detail below as a flowchart in **Figure 2.1**.

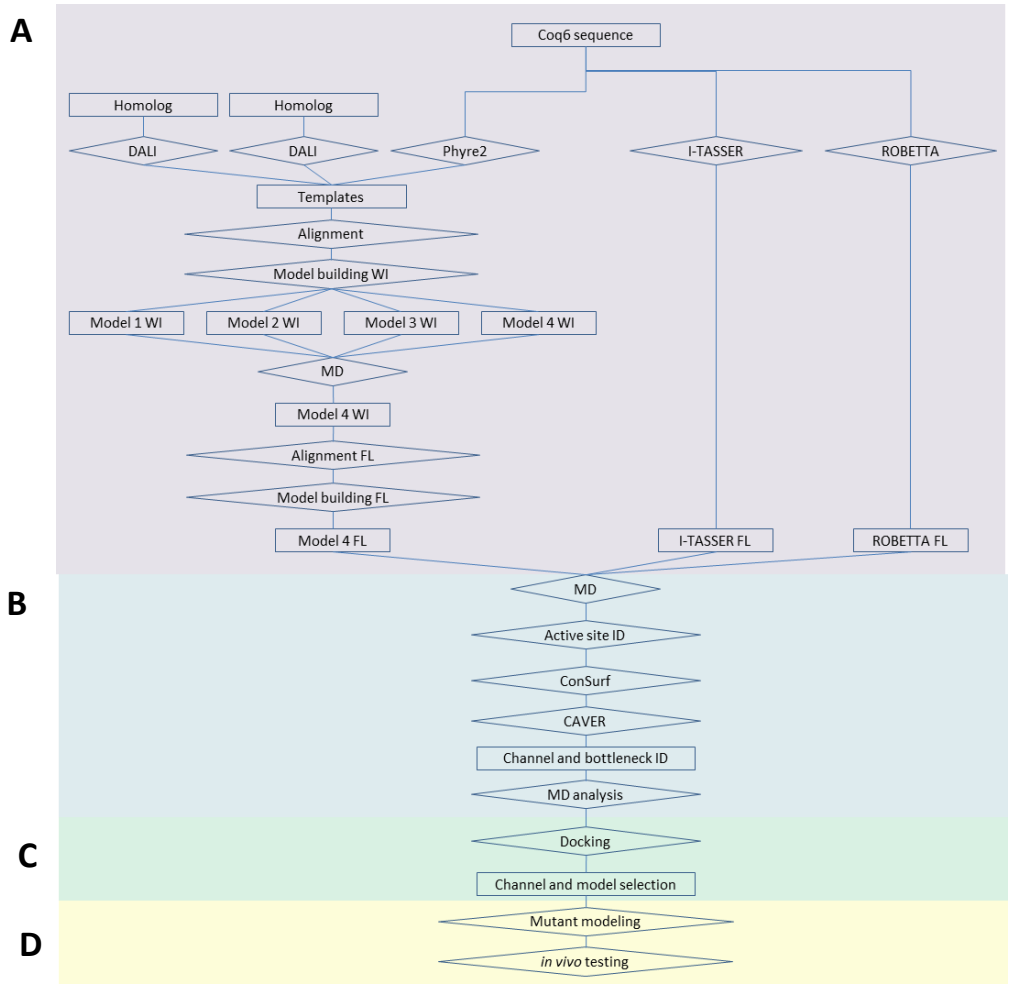

FIG 2.1: *Modeling strategy flowchart. The approach is divided into four main sections. A) Homology modeling, which consists of template search, sequence alignment, and model generation and screening. B) Evolutionary residue conservation and geometric analysis of structures from molecular dynamics simulations. C) Development of the hypothesis of a substrate binding regions followed by substrate docking. D) Design and in silico and in vivo testing of enzyme mutations to test the hypothesis developed in part C.* ***WI*** *designates Coq6 models Without Insertion,* ***FL*** *designates Full Length models including the Coq6 family insertion.*

3. Overview of homology modeling

Homology modeling is the process of constructing a molecular model of a protein of unknown structure on the basis of a similar protein of known structure. It is composed of four main steps: i) template searching, ii) sequence alignment, iii) model building. Model evaluation is subsequently performed through MD calculations, presented in Section 4. In the present section, we will describe the theoretical basis of the techniques used in each step of the production of homology models, as well as the specific software implementation we have chosen for each step.

3.1 Template searching and alignment

3.1.1 The importance of finding good templates

The general goal of homology modeling is to create a mapping between the residues of the target (Coq6 in our case) and the residues of a template (a protein of known structure). This sequence-to-sequence mapping is then used to generate a 3D model of the target protein, where the amino acids of the target are constructed at coordinates corresponding to their homologous residues in the template. In order for the mapping between the target and template residues to be plausible in 3D, the sequences of the target and template must be similar, and can be quantified by the *sequence identity*: the percentage of identical residues at the same position in the target-template alignment. The relationship between sequence identity and structural similarity (as measured by the RMSD of homologous atoms) is resumed in **Figure 2.2** below.[1]

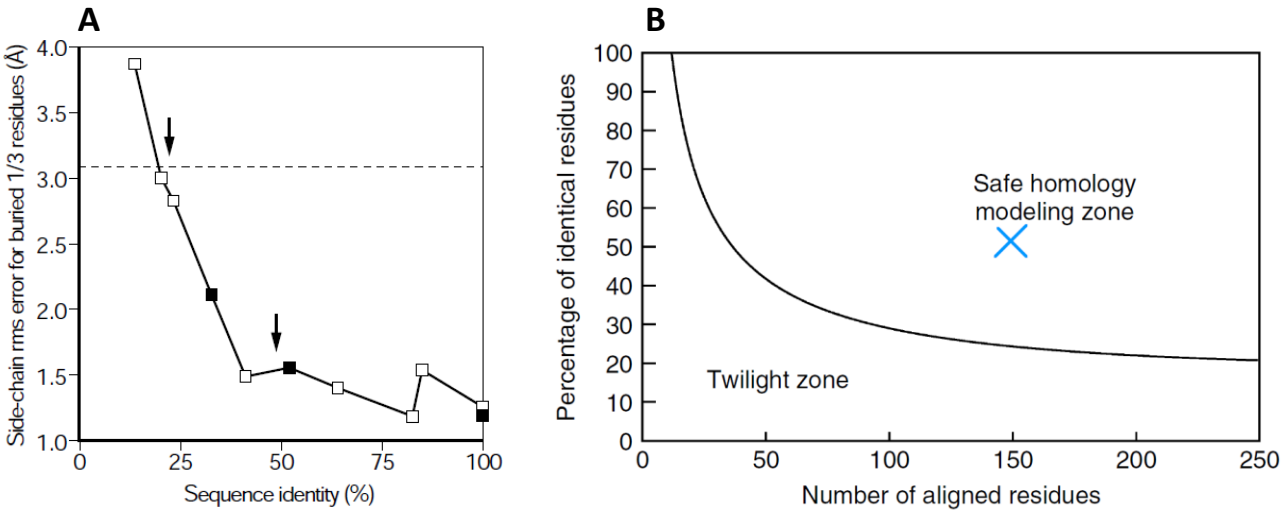

FIG 2.2: *A) The relationship between sequence identity and structural similarity as derived from a study of the PDB.[2] Structural similarity as assessed by sidechain RMS for core residues plotted as a function of sequence identity shows a large jump around 30%, defining the limit of the twilight zone. Above 30% sequence ID, two sequences are very likely to adopt the same fold. Below this limit, the sequences could have entirely different folds. B) Sequence identity plotted against sequence length[1]; the contour is derived from another study of the PDB (Rost 1999).[3]*

Above a threshold of 30% sequence ID, sequence identity is strongly correlated with structural similarity (as measured by RMSD), and the templates found for any given target are likely to have the same global fold, and are often in the same protein family. Homology modeling in this regime of sequence identity is likely to produce a model with predictive ability. When the sequence ID is below 30%, the templates found for any given target may have different global folds from each other, meaning that the sequence identity is too low to reliably detect a single general conformational "solution" (the global fold) for the target protein sequence.[3] As we can see from **Figure 2.2**, there is a very high similarity between protein structures when their sequences are at least 14% identical. However, below this limit, protein sequences can have entirely different global folds. This region of the plot is often called the twilight zone, and for good reason: just as in twilight lighting conditions, contrast is low which makes resolving objects and comparing them difficult.

For the Coq6 sequence, the best templates (discussed in more detail in Chapter 3) show a sequence identity of 15-20%, well below the threshold for creating accurate models. Therefore, it is of paramount importance to find and select the best template structures because this choice will have the largest impact on the quality of the final result obtainable. This is point is of such importance we will review the reasons behind it.

The relationship between protein sequence and structure is still incompletely understood. Protein structure is more conserved than protein sequence.[4] For example, consider three proteins A, B, and C. Protein A has a lower sequence identity to protein C than protein B, but it may still be more structurally similar to protein C than protein B. It is still not routinely possible to accurately predict the structures of known proteins using purely theoretical approaches for two reasons.

One fundamental reason for this is physical, and arises from classical approximations to quantum mechanics. We cannot simulate the effect of electronic structure on interatomic forces and molecular geometry using quantum mechanics (QM) at a practically useful speed for molecular systems as large as proteins. We must typically use classical mechanics approximations of inter-atomic interactions which must be empirically parameterized against experimental or QM-calculated data in order to reproduce their molecular geometries. These approximations allow us to simulate molecular structure and motion on much larger scales of space and time than computationally possible with QM, but they lead to a simplified representation of the underlying physical reality and cannot reproduce all of its behavior. This is discussed in more detail in Section 4.2.1. The practical result is that such simulations are not guaranteed to "correct" the errors arising from the difference between the target's *true* 3D structure (which homology modeling will try to approximate) and the best template's 3D structure. In other words, it is important to start as close as possible to the answer because current simulation methods can only correct relatively small errors in protein geometry.

The second fundamental reason for this is mathematical, and arises from the presence of multiple minima in a highly dimensional potential energy function. The classical mechanics approximations to interatomic forces and molecular geometry essentially relies on being able to calculate a potential energy value as a function of interatomic distances and angles. The potential energy functions used to describe even a single interatomic distance or angle value can display multiple minima separated by energy barriers which must be crossed in order sample other minima. Current simulation methods do

not guarantee that the atoms in such a system will be able to explore the geometries corresponding to all minima of the potential energy function. While this is generally not true for very small systems consisting of a few or a few dozen atoms, the dimension problem is multiplied greatly for systems the size of a protein. The potential energy of each single interatomic distance or angle can be thought of as a single dimensional slice of the entire systems' potential energy function. The potential energy function of the entire system depends on the relative distances and angles for all atoms in the system, and is therefore of very high dimension for a system the size of Coq6, which typically contains 100 000 atoms. The result is that the energy function which determines the molecular geometry observed in modeling and simulation contains *many local minima* in *a very high number of dimensions*.

The practical consequence is that molecular dynamics simulations of a finite length are not guaranteed to visit all local minima of the potential energy function and explore all possible molecular geometries. That is to say, the molecular systems we study are not ergodic – not all regions of the protein system's phase space are accessible through molecular simulation.

Indeed, molecular dynamics simulations of a finite number of steps are not guaranteed to escape or even appropriately sample the single local minimum (meaning a single initial molecular geometry) from which it was started. This makes it very important to start any modeling or simulation process from a variety of models that differ significantly from each other because current methods will likely not be able to escape the geometry and dynamics of the initial configuration. In the context of homology modeling, this makes template choice critical – especially in a scenario where the closest matching sequences have a sequence identity of less than 30%.

3.1.2 Sequence search by partial pairwise methods: BLAST

The simplest methods of finding templates uses exhaustive pairwise comparisons between the target sequence and potential templates in a protein sequence database (such as NCBI[5] or UniProt[6]). The comparison relies on computing sequence alignments between the target and the sequences in the database. Therefore we will review the basics of sequence alignment, which will also apply to the more detailed alignments to be performed after the initial database search has yielded its results.

There are two classes of sequence alignment methods in common use today: dynamic programming and "word" based methods.[7] Dynamic programming methods operate on the complete protein sequences and are theoretically able to find a single global optimum alignment between two sequences, or indeed, any number of sequences. However, the computational expense of this approach makes it impractical to apply to more than 8 sequences at once,[8] and a sequence database search contains many more than this. Word based methods are not guaranteed to generate optimal alignments (and therefore find the best results in a database), but they are much faster to compute, and therefore more applicable to searching large databases. The best known examples of this class of methods are FASTA[9] and BLAST[10].

As implied by the description of "word-based" methods, these methods select a series of shorter, non-overlapping sub-sequences (words) of residues from the query sequence and looks for matching words among the sequences in the database. The presence and relative positions of these "words" are used as sequence-specific features to recognize similarities between protein sequences without actually

performing a globally optimized alignment against each. However, because these methods rely on detecting conserved "words", they will only be able to find matches if they are already quite similar in sequence. These methods perform well when sequence identity between the target and potential templates in the database is high (above 30%). However, when a given database contains only lower sequence identity templates (in the range of 20-30%), word-based searches methods typically find only half of the *possible* templates. This is because the reliance on exact residue identity on relatively short words makes the search more sensitive to "noise" (or mutations incurred through evolution) than signal (identical residues). When two sequences have less than 30% global sequence identity, there is more noise than signal. That is to say, there are more differences than there are similarities, so recognizing templates by literal residue identity over short stretches is (understandably) likely to miss matches that may exist in the database.

What is needed to overcome this weakness is a way to represent each position in a pairwise alignment more generically, to make the representation and comparison of protein sequences less sensitive to their differences and more sensitive to their similarities.

As more advanced methods described in the following sections will show, Coq6 has a sequence identity of 15-20% with respect to its closest templates. We are operating in a zone of low sequence ID where structural divergence between a candidate template structure and the target's "true" structure is very likely. Therefore, we decided to turn to a class of more sensitive methods to find appropriate templates for Coq6: hidden Markov models. However, we will first need to introduce alignment scoring matrices in more detail, as performing operations on protein sequences with matrices is a common component to both the search and alignment procedures we will use in this project.

3.1.3 Sequence search by complete-sequence methods: PSSMs

A protein sequence serves to define a protein structure, and protein structure is more conserved than sequence.[4] This fundamental and interesting result of the field means that a single protein sequence is a direct but intrinsically limited way of describing a protein structure. Analysis of structural databases such as SCOP[11] and CATH[12] reveal that there are about 1300 unique folds that have been catalogued among about 65 000 experimentally solved structures in the Protein Data Bank[13] (PDB).

The structural implication of this is that a single protein sequence is always actually a member of a larger family of structural homologs. For the purposes of searching for structural homologs it becomes valuable to represent a particular protein sequence (whether it is of the target or a template) in a way that is more general than a single explicit sequence, yet specific enough to be associated to only a single class of protein shape, that is to say, limited to a single global fold.

The more general representation we are looking for can be constructed as a position specific scoring matrix[14] (PSSM), also known as a *profile*. The matrix is a very useful data structure for describing multiple protein sequences, and we will see it again. To construct this PSSM for any protein sequence, we first find a set of *close relatives* with a simpler method, such as one of the pairwise sequence searches. We then create a multiple sequence alignment (MSA) using more accurate methods for this set of closely related proteins with each sequence occupying a row, and each column containing homologous residues

at a generically numbered alignment position. The frequency of occurrence of each amino acid type at each position can be calculated from this MSA.

This allows us to create a new representation of the sequence. It has the general form of a matrix. Each row represents one of the 20 amino acids, and each column represents a numbered position in the MSA. Within each row (which corresponds to one of the 20 amino acids), the value at each column position is the frequency of occurrence of that amino acid among the sequences in the MSA. Thus we have constructed a *matrix* which describes a protein sequence as a series of residue frequency *scores at specific positions*: a position specific scoring matrix (PSSM).

For example, we can use a PSSM constructed for a set of templates to compute a compatibility score against a target sequence and evaluate the score to determine if the target really could be a member of the structural family used to construct the PSSM. Alternately, we could construct a PSSM for the target sequence as well, and compute the similarity between template PSSMs and the target PSSM.

The main limitation of this type of PSSM is the inability to include insertions or deletions among their protein sequences. Insertions and deletions are highly likely to occur during evolutionary divergence, particularly when target and template have less than 30% sequence identity. Therefore, while PSSMs give a useful framework for describing sequence variation among closely related sequences, we need a way to include insertions and deletions so we can use PSSMs to describe and detect distantly related sequences – specifically those with similar structures to Coq6.

3.1.4 Hidden Markov model methods: Phyre2[15]

We need an even more flexible way of representing protein sequences that can accommodate the insertions and deletions which occur during evolution. We can take PSSMs a step further with hidden Markov Models (HMM).[16] An HMM is a way of representing the biological evolutionary process as the operation of a *finite state machine*.[17] The finite state machine is a general conceptual construct from computer science. It is used to represent a system which can occupy only one of several *states* at a time, with each *state* having an *emission probability*, and *transition probabilities* for changing from one state to another.

This general formalism is applicable to protein sequences. Each *state* is an alignment position in the PSSM which can now accommodate two new "residue types": insertions and deletions. Each *state* (or alignment position) has a *probability of emitting* its own residue identity (which we recall is actually a frequency table over the 20 residue types, as in the PSSM), and each of these residue types has a specific *transition probability* of being followed by another amino acid type at the next *state* (or position in the sequence).

This is a much more complex and detailed way of representing protein sequences, but it is worth the effort. In addition to adding the representation of insertions and deletions, the transition probabilities allow an HMM to represent the mutational propensity of each position in the sequence beyond the sequence diversity captured in the underlying PSSM. This ability enables HMM-based searches to reliably detect structurally similar proteins with sequence identities as low as 15%. Since homology modeling of

an unknown target by definition begins without knowledge of templates, it is not known *a priori* how similar it is to anything in a protein structure database. It is not known if the target sequence will be above or below the threshold of 30% sequence identity with anything in the database. Therefore, it makes sense to use the most sensitive method first, in order to have the best chance of finding a template even in a worst-case scenario.[18] [19]

This is why we have chosen to begin the search for Coq6 templates with an HMM-based method. The specific implementation we have used in this project is the Phyre2[15] homology modeling server. The core of the Phyre2 server is a protein structure database, called a fold library, containing about 65 000 proteins of known structure and sequence. An HMM is built for each sequence in the fold library. When a target sequence of unknown structure is submitted for modeling, an HMM is constructed for it and Phyre2 searches for HMMs in the fold library that best match it. The resulting list of matching HMMs (and their parent protein structures) are then presented as results of the search, ranked by sequence identity weighted by sequence coverage. This is resumed in **Figure 2.3** below.

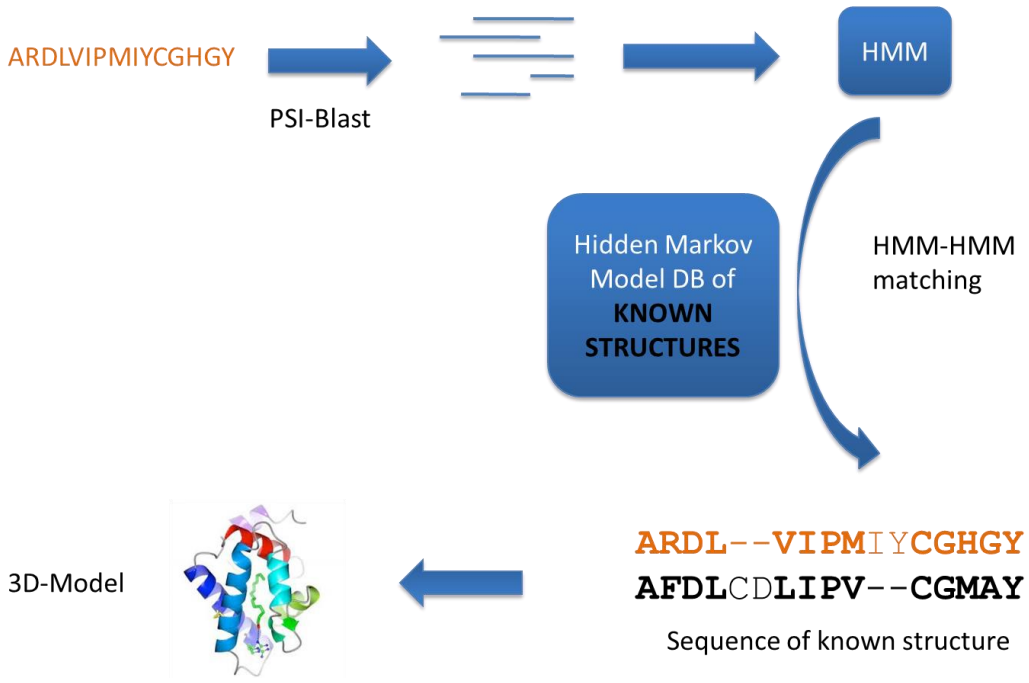

FIG 2.3: *Graphical overview of the Phyre2 workflow. Adapted from the Phyre2 website http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index*

For Coq6, this operation found templates with sequence identities in the range of 15-22%. The results (described in more detail in Chapter 3) indicated we were well below the threshold of 30%, making distinction among the top Phyre2 results ambiguous. Therefore, we turned to another method of template search, not based on sequences, but 3D structures.

3.1.5 Structure based searching: DALI[20]

Protein structure is more conserved than sequence.[4] This is one of the fundamental reasons why discriminating among the templates found by Phyre2 is difficult. When there is not a large difference in sequence identity among the templates, ranking them by more accurately refined alignments (discussed in the following section) may not lead us to the most structurally related one. Therefore, in the twilight zone, it also makes sense to expand the search by searching based on the *structures* of proteins of similar sequence, instead of by their sequences alone.

Just as for sequence based searching, structure based searching relies on a comparison of the query structure to a database of known structures. This leads to a similar challenge, which is how to accurately represent and efficiently compare structures of different proteins. Different proteins of the same length will have different side chains at a number of residue positions. This will make evaluation of protein structures by direct comparison of 3D atomic coordinates difficult, because many sidechain atoms in one structure will not be present in the others. In a calculation like the minimization of the root-mean-square deviation (RMSD) of homologous atoms, this will reduce the comparable 3D coordinates to just the backbone atoms of residues identified as homologous.

For this project we selected the DALI[20] (Distance Alignment Matrix) method, which represents individual protein structures as residue contact maps. A residue contact map is a square matrix where the rows and columns are the residues of the same proteins. Positions in this matrix contain values indicating whether a given residue forms a contact with any other residue in the protein's own sequence (and therefore its own structure). This gives a two-dimensional description of three-dimensional structure. An example is shown below in **Figure 2.4**.

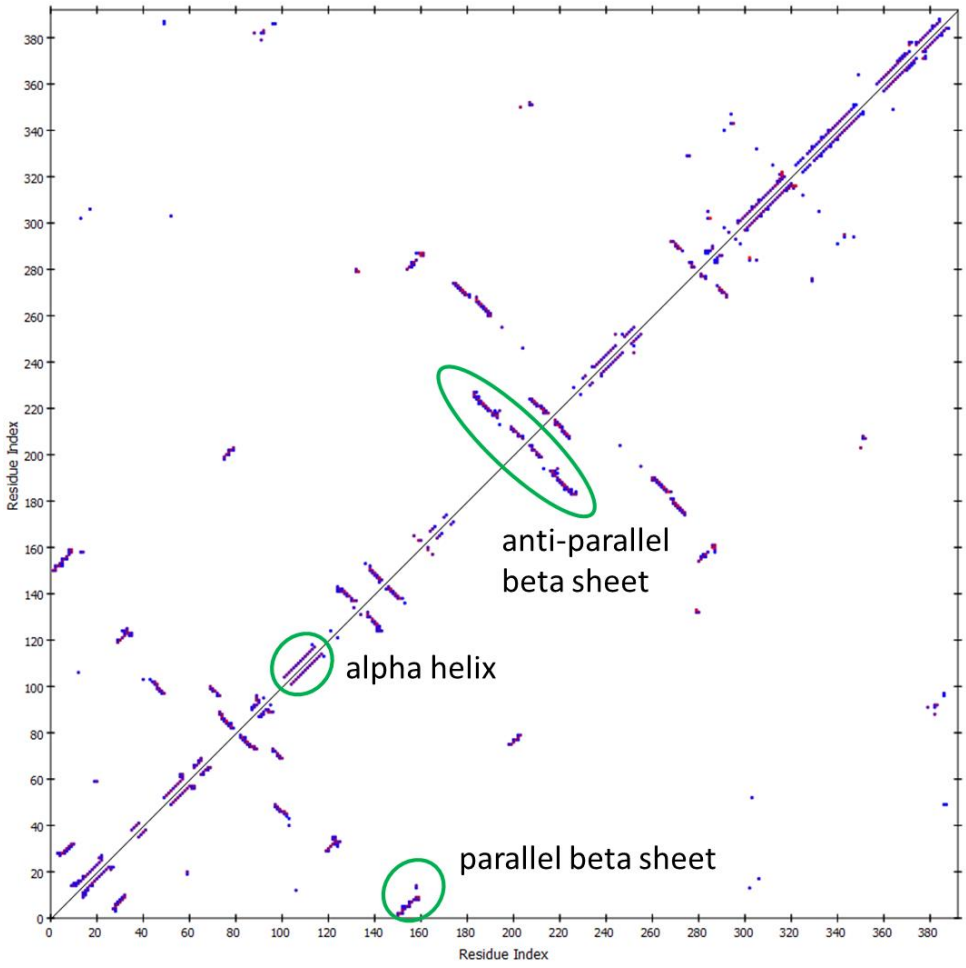

FIG 2.4: *Alpha carbon contact map for PHBH structure 1PBE[21] showing the patterns formed by the major types of secondary structure.*

The elements of protein secondary structure produce characteristic patterns on such a contact map. Alpha helices form stripes very close to and parallel to the matrix main diagonal because they consist of contacts along a stretch of *consecutive* residues. Anti-parallel beta sheets form stripes originating from and perpendicular to the main diagonal because they consist of contacts between some residues which are start off close in sequence (residues near the turn in anti-parallel beta sheets) but then progressively become more distant as the sheet extends along the sequence. Parallel beta sheets form stripes parallel to but distant from the diagonal because they necessarily contain an intervening structural motif to allow main-chain direction reversal. Therefore, contact maps are capable of representing both secondary structure (usually dominated by short-range interactions) and tertiary structure features (usually dominated by long-range interactions) in a two-dimensional data structure that does not directly depend on manipulating explicit 3D coordinates of non-homologous atom sets. This makes them accurate and efficient representations of protein structure for structural comparison. Proteins which are structurally similar will have similar features in similar locations on their respective contact maps. DALI compares the contact map calculated from a query structure to the contact maps it has already computed for all the

known structures in the PDB. Essentially, DALI divides each protein's contact map into several sub-matrices and then tries to maximize the overlap of structural features.

The matrix format of this structural representation again lends itself to match searching and is also a natural format for computing alignments, since the rows and columns of the contact map are also labeled with their residue identity. Therefore, the superposition of two protein contact maps can be used to generate a structure-based sequence alignment.

This structure based pairwise sequence alignment is used to identify homologous residues for performing 3D superposition of the two proteins by RMSD minimization. This step does require the manipulation of the 3D alpha-carbon coordinates of the proteins. However, because the two proteins have already been selected as being very structurally similar through the intermediary of the contact map, superposition converges more rapidly to a structurally meaningful result. In a final step, the RMSDs of each structure in the database matching the query structure are used to compute a Z-score for structural similarity, producing a list of PDB structures ranked by Z-score.

3.2 Sequence alignment

Once we have found templates using the HMM-based and structure-based search methods, we must create a more detailed alignment between the template and the target sequences. The low sequence identity between the target and the templates found can make alignments using automated methods ambiguous and difficult. This is the case in the present work, particularly because the Coq6 sequence is longer than any of its templates by about 80 residues. While the first 20 or so residues can be expected to form the mitochondrial signal sequence, there are about 50 residues remaining to be placed in the sequence alignment – a formidable challenge for automated methods. The low sequence identity in our pairwise target-template alignments can be described as a *low data content*. The target-template sequences are so different that we cannot rely on them alone to reliably calculate their similarity. This will require us to supplement our pairwise alignments with multiple-sequence alignments computed from Coq6, which will give us a better context for deciding which residues to align between Coq6 and any template structures.

3.2.1 Pairwise alignment methods

Despite their theoretical ability to produce globally optimal alignments on small numbers of sequences, the best dynamic programming methods, the Needlemman-Wunsch and Smith-Waterman algorithms, did not give plausible results for the template-target alignments. This is likely because of the large number of "extra" residues in the Coq6 sequence not present in any templates. Evolution commonly produces insertions and deletions in a given protein sequence, typically occurring where they can be structurally tolerated in the protein's 3D structure. However, the insertion of 50 residues between a template and a target is a challenge for even the best alignment algorithms. Therefore, we decided not to rely entirely on automated methods for our pairwise alignments.

Protein sequences within a given family are likely to be better conserved than sequences from "templates" found through various search methods. After all, structural templates can be quite distant in

sequence, as in our case, primarily because they are selected not only the basis of maximum sequence homology, but also because their structures have been solved. A multiple sequence alignment of Coq6 against the Coq6-family of sequences contains much more relevant residue frequency data and can help us understand where mutations, insertions, and deletions are tolerated within the Coq6 sequence and predicted global fold. This will inform us as to where to place insertions and deletions in our pairwise target-template alignments.

3.2.2 Multiple sequence alignment methods

There are three main types of multiple sequence alignment methods: dynamic programming, progressive methods, and iterative methods. As mentioned before, dynamic programming methods are too computationally expensive to use for more than 8 sequences. Therefore we will describe progressive alignment methods, their limitations, and their more accurate successor, the iterative methods.

3.2.3 Progressive MSA[22]: ClustalO[23]

True to their name, the progressive methods perform alignments over multiple sequences by first aligning the two most similar sequences, and then progressively aligning one additional sequence at a time until all sequences in the set have been aligned. This implies that sequence similarity must be computed in some initial manner to determine the order in which to perform the more detailed alignment. The initial calculation of sequence similarity is typically performed with an initial heuristic method, such as word-based recognition methods as described for FASTA and BLAST.

The initial rough sequence similarity calculation is used to create a guide tree, similar to a phylogenetic tree, which sorts the sequences according to their pairwise identity. The guide tree is used to define the order in which the sequences will be aligned with a more detailed method. This is done with scoring matrix representations of protein sequences. Identical residues are assigned a positive value, while mismatches are assigned lower values. Insertions of gaps in the alignment to match chemically similar residues incur a large score penalty, followed by a smaller score penalty for the extension of an existing gap. Any pair of sequences are aligned from N-terminus to C-terminus. Chemically similar residues are aligned by sliding one sequence relative to another in the matrix, with the alignment score computed at every step in the process. The algorithm attempts to increase the score computed over a given range, or window, of residues, adding gaps when necessary.

Therefore, the process will seek to maximize the alignment score serially within the sequence, from N-terminus to C-terminus, but only locally, within the range of its scoring window. This means that the algorithm can create false local optima while it is computing the alignment scores. Essentially, the alignment can be “mis-guided” by the very local nature of the alignment score computation. Additionally, as the two sequences are aligned from N-terminus to C-terminus, no changes are made to the completed N-terminal alignments, meaning that a mistake in any pairwise alignment cannot be corrected as the algorithm proceeds to the C-terminus.

Once the first two most similar sequences have been aligned, the next most similar sequence is added to the alignment matrix, using averaged scores at each aligned position. In this way, an averaged score is

developed at each position as each new sequence is progressively added to the alignment. As each new sequence is added to the MSA, it is fixed. That is to say, any errors in the initial sequence alignments cannot be corrected by new data from the next sequence being aligned. Taken together, these aspects of the purely progressive algorithm create three weak points.

First, the results of the alignment are dependent on the order in which the sequences are aligned. The order of sequence alignments is a function of the initial rapid "word-based" sequence similarity calculation used to create the guide tree which defines the alignment order. If the guide tree is not constructed in the "true" order of sequence similarity, then the first two sequences selected for matrix based alignment will not be the most similar. This, in turn, will increase the chances of the alignment of the first pair of getting stuck in a false local optimum of alignment score during the matrix based alignment. Second, any errors incurred during the sequence alignment of any single protein are not corrected within that alignment operation. This means that the sequence alignment quality usually gets worse towards the C-terminus. Third, because alignments become fixed as they are completed, errors made within any given sequence alignment cannot be corrected by the addition of new data from the next sequence. This means that sequences aligned later in the procedure (because of a lower initial ranking from the guide tree) usually have lower alignment quality than those aligned in the beginning.

Finally, there is a specific feature of the Coq6 sequence – the presence of about 50 extra residues – that is likely to severely challenge scoring functions not designed to deal with such disparity. Given the need to place about 50 gaps in the template sequence, the scoring function is likely to be biased towards creating several shorter gaps rather than a single long gap. This bias may guide the alignment towards or away from the "true" alignment, but this is not known *a priori*. In this thesis, we used the ClustalO implementation of the progressive MSA method.[23]

3.2.4 Iterative MSA: MAFFT-L-INS-I[25]

These weaknesses are addressed with the class of iterative methods. Iterative methods begin with the same steps as the progressive methods, except they use an objective function to evaluate the results of the first MSA. The first generation purely progressive MSA is then used as the basis for realigning sub-regions of sequences so as to improve the score of the objective function. This creates a second generation of the MSA, which is then further refined until convergence of the objective function is achieved.[24] This ability to modify existing alignments enables iterative methods to correct initial errors arising from guide tree errors and local scoring errors.

The specific implementation of the iterative MSA method we have chosen for computing the Coq6-family MSA is MAFFT-L-INS-I,[25] as used by the ConSurf server.[26] The results of this Coq6-MSA are presented in Chapter 3, and are used to manually curate the pairwise alignment between target and template.

3.3 Model building

Having created the template-target alignment necessary for model building, we will now review the methods of creating 3D coordinates for the atoms of the residues of the aligned sequences. Our work has relied almost entirely on the MODELLER[27] software for the creation of our models. We have also used the automated I-TASSER[28] and ROBETTA[29] servers to create models based on different templates. These servers use their own methods for template search and alignment, similar to the methods we have selected for our own protocol, described in their respective publications.

3.3.1 MODELLER[27]

A 3D protein structure can be described by the positions of each atom in an external coordinate system, typically Cartesian coordinates, as in many common 3D structure file formats, such as PDB, and XYZ. However, there is another, more flexible way of describing atomic positions with a type of internal coordinate system. This takes the form of spatial restraints, a set of distances which define the distance from one template atom to nearby template atoms. This is similar to the way protein structures are defined in NMR. The target-template alignment is used to identify homologous residues in the target. Distance restraints derived from the template structure are applied to the residues of the target structures, constraining them to a relative geometry which will recapitulate the relative geometry of the template. Additional restraints on molecular geometry building, such as bond lengths, bond angles, and dihedral angles are also incorporated through the application of a molecular mechanics force field. These knowledge-based and force-field based restraints are combined into an objective function which depends on the coordinates of all the atoms in the nascent target model. MODELLER searches for the global optimum of its objective function through a simulated annealing process, which allows the target model to retain the general structure of the template while adapting its detailed structure under the guidance of the objective function. The procedure is repeated several times, yielding an ensemble of models, typically showing more conformational variability in regions of the proteins with fewer distance restraints. This is usually observed in surface exposed loops and turns.

3.3.2 I-TASSER[28] and ROBETTA[29]

I-TASSER[28] and ROBETTA[29] have their own methods for template searching and alignment, which will not be described here. They differ primarily in their strategy for conformational exploration of their nascent 3D models. Whereas MODELLER uses simulated annealing, I-TASSER creates a structural decoy library, and ROBETTA uses an iterative process of rebuilding and refining in a highly accurate force field.

I-TASSER first builds its models at low resolution, representing side-chains as single particles at the center of mass. The conformations of unaligned regions are improved through lattice based Monte Carlo moves, whereas the coordinates of aligned regions are allowed to move continuously. Both types of structural moves are evaluated by I-TASSER's knowledge based force field. A set of many low energy models is retained and then clustered in order to select the cluster centroid as the most native-like structure. ROBETTA also uses Monte-Carlo moves to explore conformational space, but does not create a library of possible structures like I-TASSER. Instead, it relies on its force field to refine, and if necessary, rebuild the model so as to achieve a lower energy score.

4. Molecular dynamics

4.1 Molecular simulation

The goal of molecular simulation is to gain insight into the physical structures and processes present at the molecular scale, which are often difficult to observe experimentally. Understanding macroscopically observable phenomena at the microscopic (often molecular or atomic) scale is the foundation for much of the progress made in the study of thermodynamics.

4.1.1 From the macroscopic to the microscopic

One of these key advances was the formulation of the van der Waals equation.[30] This equation gives a more accurate description of the behaviors of gases and liquids over wider ranges of temperature and pressure than previously possible with the ideal gas law.[31] The van der Waals equation adds corrective terms to the ideal gas law for two of its simplifying assumptions.

First, molecules have a definite physical volume (not accounted for in the ideal gas law), which limits gas compressibility at higher pressure. Second, molecules interact when they approach at close range. These interactions are generally attractive (arising from favorably correlated fluctuations in electron distributions), which tends to reduce gas volume at higher pressures. However, below a certain distance (as imposed by higher external pressure), molecules will experience a strong repulsion. The addition of microscopic detail to the ideas gas law shows the important connection between macroscopic thermodynamics and microscopic molecular structure.

Since physically relevant volumes of matter are usually composed of large numbers of particles, a molecular description of matter also lends itself to a statistical description of the molecules. Another famous example of this connection was made by Josiah Willard Gibbs in his presentation of the "Elementary Principles in Statistical Mechanics"[32] which formally presented the relationships between entities at the molecular scale. We will briefly review them before proceeding to describe their practical implementation in this work.

4.1.2 From particles of matter to systems in phase space

A volume of matter which we wish to study is called a **system**. Since we are describing this system at the microscopic scale, we will represent the system as a large number of **particles**. These particles interact with each other but they also interact with their larger environment. For practical purposes, we only have the ability to represent and analyze a limited number of particles constituting the system of interest. We will represent the interactions with the environment in a much more general way, typically through general constraints such as temperature and pressure.

An **ensemble** is the set of all possible **configurations** of a **system** of **particles**, possibly subject to (and often defined by) general environmental constraints. A single **configuration** of a system is a description of the **position** and **momentum** of all its constituent particles. While many of these configurations may be different in their microscopic details (positions and momenta of all particles), many of these configurations are so similar that they are not easily distinguishable at the macroscopic scale. The detailed particle **configurations** of the **ensemble** can be **mapped** along the axes of selected system

parameters (selected for physical relevance), defining a **phase space** of the system. Phase space is a mathematical construct for conceptually organizing and describing the particle configurations that are physically possible. The change over time of a configuration through its phase space yields a **trajectory**. In the molecular simulations of this work we use computed approximations of the physical forces acting on particles to generate new configurations evolving from an initial configuration. The computation of these new configurations from individual particle trajectories is the field of **molecular dynamics**.

4.1.3 Algorithmic implementation of ensemble constraints

In realistic physical settings, the set of all configurations accessible to a given system is bounded by some constraints coming from the environment beyond the system. These constraints will place some limits on the behaviors possible for the particles in the system, and for the system in its phase space. These constraints are fairly intuitive to understand from a macroscopic point of view: volume (V), energy (E), temperature (T), and the number of particles (N). These can be used to define **statistical ensembles** representative of the configurations accessible to the system.

Systems of study are often in a confined area, where the volume (V) is constant. In the case where the system is perfectly isolated from its environment, the number of particles (N) cannot change, and the energy (E) of the system must be constant (as no energy can be lost to or gained from contact with a larger environment). This situation is described as an NVE ensemble, because the parameters N, V, and E are not allowed to change.

In the case where the system is being maintained at a constant pressure (P) by varying the volume as necessary, the situation is defined as an NPT ensemble. The NPT ensemble is a better representation of many intracellular biological systems, since they occur at a relatively constant pressure. The detailed energy E is allowed to change, but only within a limit described by the general temperature (T) of the system.

In the case where the system is not at all isolated from the environment, the number of particles (N) can also change, as can the energy (E). This situation is termed the uVT ensemble.

In molecular simulations of biological systems we are typically interested in processes occurring at constant temperature (represented by the NVT ensemble) and at constant pressure (the NPT ensemble). In practice, these constraints are implemented by specific algorithms. For NVT simulations, the equivalent of a thermostat is required to adjust and maintain the temperature; for NPT simulations, the equivalent of a barostat is also required.

In this work we select the Berendsen velocity-rescaled thermostat [56] for its fast and smooth first-order approach to the target equilibrium temperature and its ability to produce NVT and NPT ensembles (both of which are used in this work). This is of practical advantage for molecular simulations because it reduces the amount of simulation time necessary for the system to converge to the target temperature during equilibration phases of molecular dynamics as compared to other thermostats (such as Langevin[33], Nosé-Hoover[34], and Andersen[35] thermostats). These other methods can show higher order temperature oscillations requiring longer simulations to converge to the target temperature. Because it

is a weakly-coupled method, the velocity-rescaled version of the Berendsen thermostat is also appropriate for NPT simulations. Similarly, we select the Parinello-Rahman barostat for NPT simulations in this work because it can produce an NPT ensemble for a simulated system.

The final parameter we must control is (N), the number of particles in the system. This is typically accomplished with the technique of periodic boundary conditions. The volume of matter being simulated is defined as a simulation cell in which the particles move. Particles near the simulation cell's edge may have a vector which would cause them cross the simulation cell wall. Treating the wall as rigid and truly confining the particles to the cell can cause unphysical artifacts in particle behavior. Another solution is to allow a particle to exit the simulation cell from one face and then reintroduce it from the opposite cell face. This is called a periodic boundary condition, and is typical in protein simulations, including those in this work.

4.1.4 From cold crystals to warm bodies

We have briefly mentioned initial conditions as the starting point of simulations. In protein simulations, the minimal initial conditions are the coordinates of every atom in the protein, derived from experimental methods (such as X-ray diffraction or NMR) or theoretical methods, such as homology modeling. To the protein coordinates, we typically add a large shell of water molecules to explicitly represent the water surrounding proteins in biological environments. The simulation cell now contains all atoms we wish to study. However these are static coordinates, meaning they have no motion. In a physics-based simulation, this translates into an effective temperature of absolute zero. This is of course a poor approximation of biological systems, which typically operate around 300° Kelvin. Therefore, we need a way to add motion (as measured by energy or temperature) to our atoms in a physically realistic way.

In the context of molecular simulations, our virtual atoms do not feel the true physical forces responsible for defining their initial coordinates, such as intra- and inter-molecular forces. In the simulation they will feel **a computed approximation of physical forces, which can differ significantly from the real world forces**. Therefore, when we introduce protein coordinates to these simulated forces, we must adjust the protein and water coordinates according to the simulated forces in a process called **energy minimization**. The purpose of this process is to ensure that the virtual atoms do not feel unphysically large forces when we apply initial velocities to the system particles (sampled from the Boltzmann distribution at the target temperature).

This means we must first increase the temperature from 0 to 300 degrees Kelvin. This process is called **thermalization**. In this work we will perform equilibration in two steps: an NVT phase, followed by an NPT phase. The reasoning is as follows. The main purpose of equilibration is to allow the redistribution of energy through the system from the initial velocities assigned to each particle at the NVT equilibration simulation start. While the protein coordinates are relatively realistic, the water coordinates are not: water molecules are typically added in a regular array to fill all available volume in the simulation cell with an unphysical regularity. Therefore, the equilibration phase is most important for the water. The largest physical change approximated in equilibration is the temperature increase from 0K to 300K. This simulated process will greatly affect the structure of the water in the simulation, likely leading to

changes in volume, which is why we use an NVT simulation for the first equilibration phase, with the temperature increase managed by a thermostat (the Berendsen velocity-rescaled thermostat).

Once the water has formed a more realistic structure, we want to bring the system closer to the constraints of a biological system, which operate at constant pressure. This is why we conduct a second phase of equilibration, whose main purpose is to bring the system from a configuration computed in the NVT ensemble to a configuration computed in the NPT ensemble. Once equilibration has brought the system to the desired temperature (through the application of a thermostat) and pressure (through the application of a barostat), we can continue to the **production** phase of the simulation, also conducted in the NPT ensemble.

Before continuing with details of the practical implementation of our simulations, we will first describe the theoretical basis for physical forces computed in our molecular dynamics simulations.

4.2 Molecules: atomic structures and interatomic forces

Molecular dynamics simulates the movements of large numbers of atoms through the solution of Newton's equations of motion. The three dimensional arrangement of atoms in space is dictated by the interatomic forces acting between them. Molecular dynamics simulations consist of calculating the forces acting on each atom at a given point in time and adjusting their positions based on these forces to arrive at a new arrangement of the same atoms a short time later. This process is repeated iteratively to produce what looks like a "movie" of how the atoms move over time. The basic form of the equation used to calculate the forces is given in Equation 1 below, which is Newton's second law of motion. F is the force vector computed as the product of the particle mass and acceleration vector.

$$\vec{F} = m\vec{a} \qquad \textbf{Eq. 1}$$

The simplest such system would be a simulation of two atoms interacting in a vacuum. In such a simulation, we would only have to calculate forces between the two atoms. The forces are calculated as the second derivative of the potential energy. The potential energy is calculated as the sum of a set of terms which are functions of interatomic distances and angles. In this particular simulation, we only need to calculate the forces between a single pair of particles. The basic formulation of this is given in Equation 2. However, this same approach can be extended to much larger numbers of atoms. We will describe the potential energy function (which appears as E in Equation 2) and dynamics calculations based on it in more detail in the following sections. Equation 2 is Newton's second law of motion rewritten as the second derivative (∇) of the potential energy function of the system, E, where r_{ij} is the distance between atoms i and j.

$$\vec{F} = -\nabla \mathrm{E}\, \overrightarrow{r_{ij}} \qquad \textbf{Eq. 2}$$

Here, we will use molecular dynamics for two purposes: one will be to assess the physical stability of the homology models created in the previous step. The other will be to perform conformational sampling. A molecular dynamics simulation of a protein structure yields a *trajectory*, which is essentially a "movie"

showing the time evolution of the molecular system. Physical force functions are applied to every particle in the simulations and are iteratively recomputed at very small timesteps – on the order of 1 to 5 femtoseconds (these values are derived from the constant-acceleration assumption, discussed on page 68) , and all particle positions are recorded at a particular interval as a "snapshot" of the system after being subjected to this physics simulation. The concatenation of all these snapshots creates a *trajectory* file, which records the atomic positions at every recorded timestep, and therefore shows the movements of the atoms over the time period of the simulation. This is called a **trajectory** because it shows **one possible path of the entire molecular system through** its phase space **(the space of all possible atomic arrangements** of the system).

This definition is important because proteins visit different regions of their phase space in a semi-stochastic manner. The stochastic nature of the movement in phase space comes from the thermal motion of particles at the molecular scale of time and three-dimensional space. This thermal motion is essentially biased in certain directions by the overall structure of proteins, which display anisotropic vibrations and movements which evolution has sculpted (through selection of protein mutations) to enhance the function of each protein.

Of the many conformations possible to a protein, and of the subset of conformations visited during any given simulation trajectory, only an even smaller subset will be complementary to the binding of ligands. The same cannot be said of the substrate, which by its requirement to be released after catalysis, must undergo a more transient binding, and which therefore has a binding site which is likely to adopt a broader range of conformations, only some of which are compatible with substrate binding. Therefore, the recognition of the distinct conformational states compatible with substrate binding occurring in a trajectory is an important task, developed in Chapter 4 (Selection of Coq6 models through molecular dynamics and substrate docking).

While apo-protein MD simulations do not tell us directly about the dynamics of receptor-ligand interactions, they can still be useful for the conformational sampling of the apo-structure. This is important because the apo-protein may transiently assume a ligand-binding conformation even in the absence of a ligand, a phenomenon described in the conformational selection hypothesis of protein-ligand interaction[36] and discussed in more detail in Chapter 4 (Selection of Coq6 models through molecular dynamics and substrate docking).

4.2.1 Physics and functional representation: the potential energy function

The forces between atoms arise from the interactions of their electronic structures. This is true for both bonded and non-bonded interactions. Electronic structure is best described by quantum mechanical methods. However, these methods are too computationally expensive to apply on systems the size of proteins, which typically contain 100 000 atoms. Therefore, the effects of electronic structure on molecular geometry are described with a classical-mechanics approximation. This approximation is empirically parameterized to recapitulate the geometry predicted from quantum mechanics or observed in experimental structures.

These classical approximations have the form of potential energy functions, and consist of a set of terms constructed to represent specific constraints on molecular geometries. The potential energy function must be applied to every particle in the system, which has the potential (no pun intended) to be very computationally expensive. One way to apply the potential energy function to all system particles is to compute the function only between pairs of particles that are close enough to interact, as specified by the terms of the potential. Pairwise interactions can be calculated rapidly enough for well-defined pair lists to permit particle motion to be computed at a practical speed.

The potential energy function referred to as E in Equation 2 is composed of several terms designed to approximate the intrinsic preferences of molecular geometry arising from electronic structure effects. An example of a potential energy function is presented below in Equation 3. This shows the general functional form of the AMBER potential energy function used in molecular dynamics, adapted from Cornell et al 1995).[37]

$$E(r^N) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_a(\theta - \theta_0)^2 + \sum_{\text{torsions}} \sum_{n} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

Eq. 3

The first three terms describe the geometry of covalent bonds, while the last two describe non-covalent interactions: the van der Waals interaction between uncharged atoms and the electrostatic interaction between atomic partial charges. The first term (in the purple box) describes the stretching of a covalent bond, formulated as a harmonic approximation, where the distance of the bond is defined between two bonded atoms. The second term (in the blue box) describes the bending of a covalent bond, also formulated as a harmonic approximation, where the angle is defined between three covalently bonded atoms. The third term (in the green box) describes the torsion (or twisting) of a covalent bonds with a periodic function, where the torsion angle is defined as the angle between two planes. Each plane is formed by a unique set of three consecutive atoms from a set of four consecutive atoms. The fifth term describes interactions between atoms that are not covalently bonded with two components: the van der Waals interaction (shown in the orange box) and the electrostatic interaction between atomic partial charges (shown in the red box).

The potential shown in Equation 3 can be used to calculate the potential energy of a molecular system at a single instant in time. As shown in Equation 2, the potential energy can be used to calculate the forces acting on each atom in the system. Combining Equations 1 and 2 results in Equation 4, shown below.

$$-\frac{dE}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \qquad \textbf{Eq. 4}$$

Here we see the appearance of the term *t* (for time), derived from the acceleration term (*a*), appearing in Equation 1. Taking the simple case where acceleration is constant, we can write Equation 5. We will discuss the validity of the constant acceleration assumption below.

$$a = \frac{dv}{dt} \qquad \textbf{Eq. 5}$$

Integrating Equation 5 gives us Equation 6.

$$v = at + v_0 \qquad \textbf{Eq. 6}$$

Integrating Equation 6 again gives us Equation 7.

$$x = vt + x_0 \qquad \textbf{Eq. 7}$$

Equation 7 appears simple, but we must remember two key facts. First, in a molecular simulation, it must be applied three times (once for each spatial dimension) for every atom in the simulation (of which there can be many, often in the range of 100 000). Secondly, the velocity term (*v*) must be calculated from the acceleration, which depends on the calculation of the potential energy function (Equation 3). **Applying Equation 7 to a molecular system allows us to calculate the position of every atom at a future time (*x*) based on their position at the current time (x_0). For a molecular system of this size, the potential energy function becomes too complex to solve analytically. However, we can numerically compute an approximation to Equation 3 and its time-dependent derivatives.** Therefore, the next step is to produce a formulation of Equation 7 that can be computed accurately and rapidly. The Taylor expansion provides a tool for this, as presented in Equation 8 below.

$$\mathrm{x(t + dt)} = \mathrm{x(t)} + \mathrm{v(t)dt} + \frac{1}{2}\mathrm{a(t)dt^2} + \cdots \qquad \textbf{Eq. 8}$$

This formulation of the equation of motion contains a velocity term (*v*). In molecular simulations, the numerical values for velocity must be retained to a high precision for every atom. Including velocities in this equation makes it slower to compute. An alternative formulation for computing future atomic positions can be given by the Verlet algorithm. Recalling that the general goal is to compute future atomic positions from past atomic positions, we can write Equation 8 twice, once for the past, and once for the future. The *d*t terms have a positive sign when computing the future, and a negative sign when computing the past.

$$\mathrm{x(t - dt)} = \mathrm{x(t)} - \mathrm{v(t)dt} + \frac{1}{2}\mathrm{a(t)dt^2} + \cdots \qquad \textbf{Eq. 9}$$

Combining Equations 8 and 9 allows us to obtain Equation 10, the Verlet algorithm.

$$x(t + dt) = 2x(t) - x(t - dt) + a(a)dt^2 \qquad \textbf{Eq. 10}$$

At this point we can return to the constant acceleration assumption used to derive Equation 5. Several terms of the potential energy function prescribe harmonic motion, which gives atoms subject to it variable acceleration over the timescale of the harmonic motion. This would invalidate the constant acceleration assumption used to derive Equation 5, which contains a *d*t term. **In a discretized numerical computation of analytical forms containing a *d*t term, the *d*t term is called the time-step**. If we can make the numerical value of the time step sufficiently small, atoms will only have had time to move a very short distance, over which the acceleration cannot vary much. **By specifying a sufficiently small time step, we can satisfy the assumption of constant acceleration**. Of course, the size of the time step is relative to the natural period of the motion of the atoms. The faster the motion of the atoms, the smaller the time step must be in order to satisfy the constant acceleration condition for the use of Equation 5.

This means that the time step will be defined by the natural period of the fastest atomic motions in the molecules. The speed of atomic motion in the simulation is inversely proportional to the atoms' mass, meaning the lightest atoms will be moving the fastest. In biomolecular simulations, this corresponds to the covalent bond stretching between hydrogen and non-hydrogen atoms, which limits the timestep to 1 femtosecond. It is desirable to increase the size of the time step, in order to increase the speed of producing simulations, while maintaining physical accuracy. One way to do this is to effectively remove this fastest degree of freedom by making the hydrogen-heavy atom bonds of fixed length. In the GROMACS[49] implementation of molecular dynamics, this is done by applying a constraint to this bond length using an algorithm called LINCS[38] [39] (LINear Constraint Solver), allowing the timestep to be increased to 2 femtoseconds, effectively doubling the speed of computing the simulation.

So far we have referred to covalent bonds in a generic way. However, when we are simulating specific molecules, we will have to specify the atoms involved in each specific bonded or non-bonded interaction. That is to say, we will also have to define the atom type (which defines mass) as well as the bond type (e.g. single, double, triple) as they occur in the molecule. For each combination of atoms and bonds in a molecule of a given net charge, we must also assign a partial atomic charge to each atom. The partial charge of each atom in a molecule depends not only on its composition but also its conformation, as the molecule's conformation is fundamentally interrelated with its electronic structure. However, this interrelationship must be calculated with quantum mechanical methods, which are typically too computationally expensive to apply to the number of atoms (about 100 000) in common molecular dynamics simulations. As an approximation, a fixed partial charge is assigned to the atoms in a given molecule This parameter (as well as the other parameters appearing in the terms in **Equation 3**) are calculated using quantum mechanical calculations, in a process called parameterization. Parameters are calculated with quantum mechanical methods, an example of which is RESP[40] (Restrained ElectroStatic Potential). This method calculates the electrostatic potential over a given molecule using quantum mechanics and then assigns a partial charge to each atom in the molecule. The restraints are used to help decrease the partial charges assigned to more buried carbon atoms.

Each molecule must have its own parameter set in order to describe its intrinsic and extrinsic geometric preferences; this parameterization must be empirically or semi-empirically created for each molecule type we wish to simulate. In the case of protein modeling, the modular nature of proteins means that we can only have to develop parameters for each of the amino acids, allowing us to simulate any protein. However, other molecules we might need to include in the simulation (such as water, lipids, and ligands) may not be as common as the amino acids, and must be parameterized for the relevant force-field before being included in molecular simulation.

While we have described the terms of the potential in a general way, specific implementations differ in the functions they use for describing each of the force terms and combining them, as well as in the atomic partial charges and basic atom types. **Specific implementations of each potential, called force-fields**, are parameterized in their term values so as to reproduce molecular geometries observed in experiment and QM calculation. Many force-fields have been developed, and each is typically specialized for a specific class of molecules, such as polymers, liquids, or crystalline materials. The parameters developed for one force-field will not be valid for another because of their differing choices of functions for each term and partial charge assignment, as well as their different atom-types. Some combine non-polar hydrogens with their parent atoms in united particles (such as GROMOS[41]), while others represent all hydrogens explicitly, such as CHARMM[42], AMBER[43] , OPLS-AA[44]. Indeed, simulation results of the same molecule under different force-fields may not be directly comparable because **each force-field is representing physical reality in a slightly different way**.

Indeed, it should be repeated that all of these force-fields are empirical classical approximations to a fundamentally quantum system, and that these approximations of the real-world potential energy function are not perfect. They each have differently biased representations of the physical forces in the real world. The only way to test the quality of these approximations is to run and analyze molecular dynamics simulations and see how well they can reproduce the behaviors of known systems.

The ability of these force-fields to reproduce atomic positions within a molecule is variable. This is particularly important for the simulation of protein structure, both in model construction and molecular dynamics. For the case of producing static protein structures, force-field based methods may be of some assistance in initial stages of model building and refinement. However, subjecting the atomic coordinates of a protein to a force field in energy minimization for too long can actually cause the structure to deviate from the experimental structure.[1] This practical observation shows the limits of one of the central hypotheses of molecular modeling of proteins: the native conformation of a protein is the lowest energy conformation. This may be conceptually true, but computing the energy correctly in practice is not guaranteed to produce this result. This discrepancy between the real-world potential energy function and our numerically approximated potential energy function can also distort the results of molecular dynamics, which is something we must consider in our choice of force field.

4.2.2 Force-field selection: AMBER99-SB-ILDN[45]

We chose the AMBER99-SB[45] force field for simulating the Coq6 system for several reasons. According to a benchmark study comparing the structural evolution of proteins as simulated under several force fields compared to NMR data, different force fields favor different secondary structure types in comparative

MD simulations.[46] The study concluded that many currently used force fields (OPLS-AA, CHARMM22, GROMOS96-53a6, AMBER99-SB, and AMBER03) do not properly reproduce structural features of hydrogen bonds, which leads to unrealistic protein conformations over the course of longer simulations lasting hundreds of nanoseconds. The only force field for which this behavior was not observed was AMBER99-SB, suggesting it is a robust choice for our simulations. AMBER99-SB also has several membrane lipids already parameterized, making it a good choice for eventually simulating membranes, which may be relevant for Coq proteins. This force-field also has a relatively straight-forward procedure for parameterizing new molecules we may want to add to our dynamic simulations, such as Q biosynthesis intermediates. The AMBER99-SB force-field was further improved by comparison to additional NMR data and QM calculations[47] for the residues isoleucine, leucine, aspartate, and asparagine, yielding AMBER99-SB-ILDN as our choice for modeling Coq6. The basic AMBER99-SB force field was developed using the Restrained Electrostatic Potential (RESP)[48] model to determine atomic partial charges and parameters.

4.3 Molecular dynamics simulation code: GROMACS[49]

Each force field is a set of functions describing the potential energy of particles we wish to simulate and tabulated values for each particle type. To actually run a simulation, we need to numerically compute these functions, and this is the role of specific simulation software. Several software implementations of the force-fields are available to choose from. Molecular dynamics can be computationally expensive because of the need to evaluate many force terms for a large number of atoms (about 100 000 for the Coq6 system) at a very small time-step. Therefore, it is important to have high performance simulation codes which can perform these calculations accurately and rapidly. For this reason we have chosen GROMACS[50] (version 4.6.5), which is one of the fastest molecular dynamics software packages and widely used in biomolecular simulation. The selection of GROMACS is based on the combination of its computational efficiency as well as its compatibility with new computing hardware developed specifically to accelerate numerically intensive simulations. This aspect is described in the section on Computing Resources.

4.4 Molecular dynamics protocols

The Coq6 protein models were solvated using TIP3P[51] water under the AMBER99SB-ILDN[47] force field in GROMACS 4.6.5.[50] Electrostatics were treated with the PME[52] method (Particle Mesh Electrostatics) with a shifted potential for long range interactions using the Verlet cut-off scheme[53] set to 10Å. FAD parameters were taken from a study of flavoproteins by Sengupta *et al*.[54] The salt concentration was set to 0.157 M NaCl as documented for the mitochondrial matrix.[55] The simulation cell was a rhombic dodecahedron allowing 1.4 nanometers between the protein and the box edge. Models were subjected to 300 000 steps of steepest descent minimization. Equilibration was conducted in two phases (NVT and NPT) of 250ps each at a time-step of 1fs with position restraints on protein heavy atoms using the velocity-rescaled Berendsen thermostat[56] at 300K. NPT equilibration used the Parrinello-Rahman barostat.[57] Bond lengths were not constrained during equilibration.

Production simulations for structural stability screening were run for 20ns using a time-step of 2fs, the Verlet cut-off scheme and LINCS constraints for heavy atom-hydrogen bonds. The temperatures of the

protein and solvent were coupled separately to the Berendsen thermostat with a relaxation time of 0.1 ps at 300K.

The resulting trajectories were analyzed for structural stability. The modular construction of our models (including the insert region, further described in Chapter 3) motivated us to analyze the stability of specific sub-regions of the models. VMD[58] was used to calculate the RMSD for these regions and secondary structure persistence was calculated for selected secondary structure elements.

5. Accessible volume calculation

The goal of our modeling of Coq6 is to generate structure-function hypotheses about substrate binding. Therefore, we need to identify substrate binding sites. According to a hypothesis we develop in greater detail in Chapter 4, we posit that Coq6 has a large substrate binding site that is likely to be a cavity near the active site, situated near the FAD isoalloxazine. Therefore, we need a way to compute accessible volumes within protein structures.

5.1 Voronoi meshes: CAVER

In the computer representations used for modeling and simulation, a protein structure is a set of points placed at the positions of its atomic nuclei. Each atom has a physical extent of volume, which is not described not by a single specific radius, but by its equilibrium distance to other atoms in protein structure. Some of these distances are small, and correspond to covalent bonds. The distance, and volume, between the atoms in this configuration is too small to permit the passage of any other atom. Some of these distances are larger, and correspond to non-bonded interactions, which can still be close enough to consider two atoms as touching – still too close to permit the passage of another atom. However, a molecular structure can hold two atoms in stable positions beyond this distance without there being any appreciable interaction between them. This general configuration can have a larger distance between the two reference atoms with nothing in between: a void volume, into which another atom may be inserted or pass through.

The detection of such accessible voids is the goal of this computation. The description of protein structure as a set of simple points in space lends itself naturally to a type of plane geometry analysis called the Voronoi diagram. A Voronoi diagram is method of partitioning a plane into sub-regions based on the distances between a set of points on the plane. We can compute a perimeter around each point containing a sub-region of the plane that is closer to this region than any other point in the set. A trivial case of a perfectly regular distribution of points on a grid would produce a set of regular hexagonal sub-regions, called cells, similar to the cells of a honeycomb. Voronoi diagrams allow the calculation of such cells for irregular distributions of points, and thus have broad applicability to modeling real-world structures.

This can be extended to three dimensions, where each cell becomes a volume of space. As applied to protein modeling, the set of points are the coordinates of the atomic nuclei. Lines are drawn between neighboring points for all points in the set. The midpoint of each line is determined, and then these midpoints are themselves connected by a second set of lines. This second set of line defines the

perimeters of the Voronoi cells of the system. Recalling that each point is actually an atom with a certain volume, we need to remove their volumes from the final definition of the Voronoi cells, which identify voids in the protein. Contiguous Voronoi cells can form a pathway, which can define a tunnel or channel in a protein structure.

There are several software implementations of this general idea. In this thesis we selected the CAVER 3.0 software[59] because of its robust and rapid tunnel detection algorithm based on accurate representation of atomic volumes. Considering the fact that we will obtain our protein coordinates from molecular dynamics trajectories consisting of 10 000 structures each, a precise description of atomic volumes is necessary to ensure that we screen our trajectories accurately and efficiently. This will be very important for substrate docking, as we will see in the following section.

6. Docking

Molecular modeling of enzyme-substrate interactions requires the definition of spatial coordinates for the enzyme and the substrate. At this point in our methods review we have used homology modeling and molecular dynamics to generate enzyme coordinates. Now we need a method for generating substrate coordinates in a manner consistent with the enzyme's molecular geometry. We can identify a substrate binding site through the calculation of accessible volumes. The next step is to simulate the binding of this substrate to the binding site. This is the domain of molecular docking.

6.1 Representing binding through docking simulations: AutoDock VINA[60]

Docking aims to predict the specific conformation of the substrate in the enzyme's binding site. Docking can be generally described in two steps: a search algorithm generates possible structures of the substrate in the binding site, and a scoring function (whose implementation in Vina was derived from the X-score[61] scoring function) evaluates the physical plausibility of this conformation. While the coordinates of the enzyme are fixed, the coordinates of the substrates can vary, primarily through the rotation of single bonds. Hexyaprenylated Q biosynthesis intermediates can have up to 20 rotatable bonds, most of which are in the hexaprenyl tail. This generates a large space of possible conformations, which is important to efficiently explore.

Again, this is an application where the speed (as well as accuracy) of a software implementation becomes a critical factor in using it. AutoDock is an accurate[60] and commonly used program. The protein structure is spatially discretized into a grid form which stores energy potentials for substrate interactions in an atom-type dependent manner. This pre-calculation of interaction energies onto a discretized representation makes computation more efficient because non-bonded pair interaction lists do not need to be re-calculated. Instead, the substrate is effectively docked into regions of space of varying compatibility which have been mapped over the protein structure. The latest version of AutoDock, AutoDock Vina[60], calculates these grids automatically. The large conformational space available to flexible ligands (like Q biosynthesis intermediates) searched by a genetic algorithm which does not explore the conformational space exhaustively, but is nonetheless capable of reproducing bound substrate poses for many co-crystal complexes.[60] This search algorithm makes Vina much faster than

previous versions of AutoDock, and allows us to perform a more in-depth docking characterization of our system, including the full length hexaprenyl tail of a substrate model.

7. Computing resources

The molecular modeling techniques used in this thesis can be computationally intensive, particularly molecular dynamics and therefore required access to suitable computing resources as detailed in **Table 2.2** below.

TABLE 2.2 *Computational resources used in this project*

Technique	Software	Hardware support
Homology modeling		
Template search	Phyre2[14]	on-line server http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
	DALI[19]	on-line server http://ekhidna.biocenter.helsinki.fi/dali_server
Sequence alignment	Discovery Studio 3.1	Xenon workstation
	Phyre2[14]	on-line server
	ClustalO[22]	on-line server http://www.ebi.ac.uk/Tools/msa/clustalo/
	ConSurf[25]	on-line server http://consurf.tau.ac.il/
Model building	Modeller 9v8 in DS 3.1	Xenon workstation
Homology model analysis		
Molecular dynamics	GROMACS 4.6.5	Tesla workstation
Active site identification	PDB[13]	on-line server http://www.rcsb.org/pdb/home/home.do
Accessible volume calculation	CAVER for PyMol	Tesla workstation
Evolutionary residue conservation	ConSurf[25]	on-line server
Substrate docking		
Molecular coordinate preparation	AutoDock Tools	Tesla workstation
Substrate docking	AutoDock Vina	Tesla workstation

The Tesla workstation is a Dell T7600 with 2 Intel Xeon E5-2630 processors (six cores per processor at 2.3 GHz) and 64 GB of RAM. It also has an Nvidia Tesla K2075 GPU, which is a graphics processing unit (GPU) designed for intensive numerical simulations. The primary use of this workstation is molecular dynamics (using GROMACS) and molecular docking (using Autodock Vina). The use of GPUs in scientific simulations is a rapidly expanding field, stimulated by the capacity of GPU computing architectures to deal with the large volume of parallel computations required in molecular simulations, particularly molecular dynamics. This is manifested as a co-evolution of the hardware (GPUs) and software (molecular dynamics codes) in the molecular simulation field to create high performance workstation-class computing solutions which are much more affordable and accessible than CPU based supercomputing centers.

For this project we have chosen to use GROMACS for molecular dynamics because of its speed at computing simulations and its growing ability to use GPU hardware, such as the Tesla line of NVIDIA GPUs. The combination of GROMACS and GPU computing resources gives us the ability to simulate proteins on larger scales of time and space than normally possible on workstation-class computers. The Tesla workstation built for this project is capable of computing about 10 nanoseconds per day for a system of about 100,000 atoms using GROMACS version 4.6.5. This is important because obtaining representative conformational sampling through molecular dynamics is a non-trivial task potentially requiring a large amount of simulation time.

The Xenon workstation is a Dell T7500 with 2 Intel Xeon X5647 processors (four cores per processor at 2.93 GHz) and 48 GB of RAM. The primary application of this workstation is the DiscoveryStudio 3.1 molecular modeling suite from Accelrys. DiscoveryStudio integrates many common molecular modeling tasks and calculations into a single graphical environment and transparently handles file formatting issues as data is passed from one DiscoveryStudio simulation module to another. Most importantly, it integrates the MODELLER program for generating homology models, which are the basis of this thesis. It also contains modules for performing biophysical calculations on protein structures as well as sequence analysis.

These are significant advantages, because while open-source alternatives do exist, they are typically specific to one task and must be individually acquired (some are freely available, others only through licensing, even if academic) and installed. Since each open-source software often has different prerequisite software requirements (such as compilers, libraries, drivers, etc.), this often creates a complicated software dependency tree which must be actively managed, adding significantly to the researcher's time spent on "IT overhead" tasks. This reduces the time available for actually performing and analyzing simulations. The second major drawback of using a library of open source programs is that each typically has its own particular file format, requiring conversion between each modeling task. This also adds significantly to IT overhead.

References

[1] Krieger, Elmar, Sander B. Nabuurs, and Gert Vriend. “Homology Modeling.” In *Structural Bioinformatics*, edited by Philip E. Bourne and Helge Weissig, 509–23. John Wiley & Sons, Inc., 2003. http://onlinelibrary.wiley.com/doi/10.1002/0471721204.ch25/summary.

[2] Chung, Su Yun, and S Subbiah. “A Structural Explanation for the Twilight Zone of Protein Sequence Homology.” *Structure* 4, no. 10 (October 15, 1996): 1123–27. doi:10.1016/S0969-2126(96)00119-0.

[3] Rost, Burkhard. “Twilight Zone of Protein Sequence Alignments.” *Protein Engineering* 12, no. 2 (February 1, 1999): 85–94. doi:10.1093/protein/12.2.85.

[4] Illergård, Kristoffer, David H. Ardell, and Arne Elofsson. “Structure Is Three to Ten Times More Conserved than Sequence--a Study of Structural Response in Protein Cores.” *Proteins* 77, no. 3 (November 15, 2009): 499–508. doi:10.1002/prot.22458.

[5] Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. “NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins.” *Nucleic Acids Research* 33, no. Database issue (January 1, 2005): D501–4. doi:10.1093/nar/gki025.

[6] Magrane, Michele, and UniProt Consortium. “UniProt Knowledgebase: A Hub of Integrated Protein Data.” *Database* 2011 (January 1, 2011): bar009. doi:10.1093/database/bar009.

[7] Gollery, M. “Bioinformatics: Sequence and Genome Analysis, 2nd Ed. David W. Mount. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2004, 692 Pp., $75.00, Paperback. ISBN 0-87969-712-1.” *Clinical Chemistry* 51, no. 11 (November 1, 2005): 2219–2219. doi:10.1373/clinchem.2005.053850.

[8] Lipman, D J, S F Altschul, and J D Kececioglu. “A Tool for Multiple Sequence Alignment.” *Proceedings of the National Academy of Sciences of the United States of America* 86, no. 12 (June 1989): 4412–15.

[9] Pearson, William R. “Rapid and Sensitive Sequence Comparison with FASTP and FASTA.” *METHODS IN ENZYMOLOGY*, Methods in Enzymology, 183 (1990): 63–98. doi:10.1016/0076-6879(90)83007-V.

[10] Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman. “Basic local alignment search tool.” *Journal of Molecular Biology* 215, no. 3 (October 1990): 403-410

[11] Murzin, Alexey G., Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. “SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures.” *Journal of Molecular Biology* 247, no. 4 (April 7, 1995): 536–40. doi:10.1016/S0022-2836(05)80134-2.

[12] Sillitoe, Ian, Tony E. Lewis, Alison Cuff, Sayoni Das, Paul Ashford, Natalie L. Dawson, Nicholas Furnham, et al. “CATH: Comprehensive Structural and Functional Annotations for Genome Sequences.” *Nucleic Acids Research* 43, no. D1 (January 28, 2015): D376–81. doi:10.1093/nar/gku947.

[13] Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. "The Protein Data Bank." *Nucleic Acids Research* 28, no. 1 (January 1, 2000): 235–42. doi:10.1093/nar/28.1.235.

[14] Stormo, G D, T D Schneider, L Gold, and A Ehrenfeucht. "Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in E. Coli." *Nucleic Acids Research* 10, no. 9 (May 11, 1982): 2997–3011.

[15] Lawrence, Lawrence A. Kelley, and Michael J. E. Sternberg. "Protein Structure Prediction on the Web: A Case Study Using the Phyre Server." *Nature Protocols* 4, no. 3 (February 2009): 363–71. doi:10.1038/nprot.2009.2.

[16] Karplus, K., C. Barrett, and R. Hughey. "Hidden Markov Models for Detecting Remote Protein Homologies." *Bioinformatics* 14, no. 10 (January 1, 1998): 846–56. doi:10.1093/bioinformatics/14.10.846.

[17] Cassandras, Christos G., and Stéphane Lafortune. *Introduction to Discrete Event Systems*. Springer Science & Business Media, 1999.

[18] Riccardo, Riccardo M. Bennett Lovsey, Alex D. Herbert, Lawrence A. Kelley, and Michael J. E. Sternberg. "Exploring the Extremes of Sequence/structure Space with Ensemble Fold Recognition in the Program Phyre." *Proteins* 70, no. 3 (February 15, 2008): 611–25. doi:10.1002/prot.21688.

[19] Söding, Johannes. "Protein Homology Detection by HMM–HMM Comparison." *Bioinformatics* 21, no. 7 (April 1, 2005): 951–60. doi:10.1093/bioinformatics/bti125.

[20] Holm, Liisa, and Päivi Rosenström. "Dali Server: Conservation Mapping in 3D." *Nucleic Acids Research* 38, no. suppl 2 (July 1, 2010): W545–49. doi:10.1093/nar/gkq366.

[21] Schreuder, Herman A., Peter A. J. Prick, Rik K. Wierenga, Gerrit Vriend, Keith S. Wilson, Wim G. J. Hol, and Jan Drenth. "Crystal Structure of the P-Hydroxybenzoate Hydroxylase-Substrate Complex Refined at 1.9 Å Resolution: Analysis of the Enzyme-Substrate and Enzyme-Product Complexes." *Journal of Molecular Biology* 208, no. 4 (August 20, 1989): 679–96. doi:10.1016/0022-2836(89)90158-7.

[22] Bujnicki, Janusz. *Prediction of Protein Structures, Functions, and Interactions*. John Wiley & Sons, 2008.

[23] Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. "Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7, no. 1 (January 1, 2011): 539. doi:10.1038/msb.2011.75.

[24] Hirosawa, M. Totoki, Y, Hoshida M. "Comprehensive study on iterative algorithms of multiple sequence alignment." *Computer Applications in the Biosciences* 11, no. 1 (February 1995):13-18

[25] Katoh, Kazutaka, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. "MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment." *Nucleic Acids Research* 33, no. 2 (January 1, 2005): 511–18. doi:10.1093/nar/gki198.

[26] Celniker, Gershon, Guy Nimrod, Haim Ashkenazy, Fabian Glaser, Eric Martz, Itay Mayrose, Tal Pupko, and Nir Ben-Tal. "ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function." *Israel Journal of Chemistry* 53, no. 3–4 (April 1, 2013): 199–206. doi:10.1002/ijch.201200096.

[27] Šali, Andrej, and Tom L. Blundell. "Comparative Protein Modelling by Satisfaction of Spatial Restraints." *Journal of Molecular Biology* 234, no. 3 (December 5, 1993): 779–815. doi:10.1006/jmbi.1993.1626.

[28] Zhang, Yang. "I-TASSER Server for Protein 3D Structure Prediction." *BMC Bioinformatics* 9, no. 1 (January 23, 2008): 40. doi:10.1186/1471-2105-9-40.

[29] Kim, David E., Dylan Chivian, and David Baker. "Protein Structure Prediction and Analysis Using the Robetta Server." *Nucleic Acids Research* 32, no. Web Server issue (July 1, 2004): W526–31. doi:10.1093/nar/gkh468.

[30] "Reprint of: The Equation of State for Gases and Liquids." *The Journal of Supercritical Fluids*, 100th year Anniversary of van der Waals' Nobel Lecture, 55, no. 2 (December 2010): 403–14. doi:10.1016/j.supflu.2010.11.001.

[31] École polytechnique (Palaiseau, Essonne) Auteur du texte. "Journal de l'École Polytechnique / Publié Par Le Conseil D'instruction de Cet établissement." Issue. *Gallica*, 1834. http://gallica.bnf.fr/ark:/12148/bpt6k4336791.

[32] Josiah Willard Gibbs. *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the ...* C. Scribner's sons; [etc ., etc.], 1902. http://archive.org/details/elementaryprinc00gibbgoog.

[33] Grest, Gary S., and Kurt Kremer. "Molecular Dynamics Simulation for Polymers in the Presence of a Heat Bath." *Physical Review A* 33, no. 5 (May 1, 1986): 3628–31. doi:10.1103/PhysRevA.33.3628.

[34] Hoover, William G. "Canonical Dynamics: Equilibrium Phase-Space Distributions." *Physical Review A* 31, no. 3 (March 1, 1985): 1695–97. doi:10.1103/PhysRevA.31.1695.

[35] Andersen, Hans C. "Molecular Dynamics Simulations at Constant Pressure And/or Temperature." *The Journal of Chemical Physics* 72, no. 4 (February 15, 1980): 2384–93. doi:10.1063/1.439486.

[36] Koshland, D. E. "Application of a Theory of Enzyme Specificity to Protein Synthesis." *Proceedings of the National Academy of Sciences of the United States of America* 44, no. 2 (February 15, 1958): 98–104.

[37] Cornell, Wendy D., Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules J. Am. Chem. Soc. 1995, 117, 5179−5197." *Journal of the American Chemical Society* 118, no. 9 (January 1, 1996): 2309–2309. doi:10.1021/ja955032e.

[38] Hess, Berk, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. "LINCS: A Linear Constraint Solver for Molecular Simulations." *Journal of Computational Chemistry* 18, no. 12 (September 1, 1997): 1463–72. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[39] Hess, Berk. "P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation." *Journal of Chemical Theory and Computation* 4, no. 1 (January 2008): 116–22. doi:10.1021/ct700200b.

[40] Burger, Steven K., Jeremy Schofield, and Paul W. Ayers. "Quantum Mechanics/Molecular Mechanics Restrained Electrostatic Potential Fitting." *The Journal of Physical Chemistry B* 117, no. 48 (December 5, 2013): 14960–66. doi:10.1021/jp409568h.

[41] Scott, Walter R. P., Philippe H. Hünenberger, Ilario G. Tironi, Alan E. Mark, Salomon R. Billeter, Jens Fennen, Andrew E. Torda, Thomas Huber, Peter Krüger, and Wilfred F. van Gunsteren. "The GROMOS Biomolecular Simulation Program Package." *The Journal of Physical Chemistry A* 103, no. 19 (May 1, 1999): 3596–3607. doi:10.1021/jp984217f.

[42] Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, et al. "CHARMM: The Biomolecular Simulation Program." *Journal of Computational Chemistry* 30, no. 10 (July 30, 2009): 1545–1614. doi:10.1002/jcc.21287.

[43] Hornak, Viktor, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. "Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters." *Proteins: Structure, Function, and Bioinformatics* 65, no. 3 (November 15, 2006): 712–25. doi:10.1002/prot.21123.

[44] Jorgensen, William L., David S. Maxwell, and Julian Tirado-Rives. "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids." *Journal of the American Chemical Society* 118, no. 45 (January 1, 1996): 11225–36. doi:10.1021/ja9621760.

[45] Lindorff-Larsen, Kresten, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. "Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field." *Proteins* 78, no. 8 (June 2010): 1950–58. doi:10.1002/prot.22711.

[46] Lange, Oliver F., David van der Spoel, and Bert L. de Groot. "Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data." *Biophysical Journal* 99, no. 2 (July 21, 2010): 647–55. doi:10.1016/j.bpj.2010.04.062.

[47] Lindorff-Larsen, Kresten, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. "Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field." *Proteins* 78, no. 8 (June 2010): 1950–58. doi:10.1002/prot.22711.

[48] Wang, Junmei, Piotr Cieplak, and Peter A. Kollman. "How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules?"

Journal of Computational Chemistry 21, no. 12 (September 1, 2000): 1049–74. doi:10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F.

[49] Van Der Spoel, David, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. "GROMACS: Fast, Flexible, and Free." *Journal of Computational Chemistry* 26, no. 16 (December 1, 2005): 1701–18. doi:10.1002/jcc.20291.

[50] Van Der Spoel, David, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. "GROMACS: Fast, Flexible, and Free." *Journal of Computational Chemistry* 26, no. 16 (December 1, 2005): 1701–18. doi:10.1002/jcc.20291.

[51] Jorgensen, William L., Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. "Comparison of Simple Potential Functions for Simulating Liquid Water." *The Journal of Chemical Physics* 79, no. 2 (July 15, 1983): 926–35. doi:10.1063/1.445869.

[52] Essmann, Ulrich, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. "A Smooth Particle Mesh Ewald Method." *The Journal of Chemical Physics* 103, no. 19 (November 15, 1995): 8577–93. doi:10.1063/1.470117.

[53] Páll, Szilárd, and Berk Hess. "A Flexible Algorithm for Calculating Pair Interactions on SIMD Architectures." *Computer Physics Communications* 184, no. 12 (December 2013): 2641–50. doi:10.1016/j.cpc.2013.06.003.

[54] Sengupta, Abhigyan, Wilbee D. Sasikala, Arnab Mukherjee, and Partha Hazra. "Comparative Study of Flavins Binding with Human Serum Albumin: A Fluorometric, Thermodynamic, and Molecular Dynamics Approach." *ChemPhysChem* 13, no. 8 (June 4, 2012): 2142–53. doi:10.1002/cphc.201200044.

[55] Ballantyne, Js, and Cd Moyes. "The Effects of Salinity Acclimation on the Osmotic Properties of Mitochondria from the Gill of Crassostrea-Virginica." *Journal of Experimental Biology* 133 (November 1987): 449–56.

[56] Bussi, Giovanni, Davide Donadio, and Michele Parrinello. "Canonical Sampling through Velocity Rescaling." *The Journal of Chemical Physics* 126, no. 1 (January 7, 2007): 014101. doi:10.1063/1.2408420.

[57] Parrinello, M., and A. Rahman. "Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method." *Journal of Applied Physics* 52, no. 12 (December 1, 1981): 7182–90. doi:10.1063/1.328693.

[58] Humphrey, William, Andrew Dalke, and Klaus Schulten. "VMD: Visual Molecular Dynamics." *Journal of Molecular Graphics* 14, no. 1 (February 1996): 33–38. doi:10.1016/0263-7855(96)00018-5.

[59] Chovancova, Eva, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, et al. "CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures." *Plos Computational Biology* 8, no. 10 (October 2012): e1002708. doi:10.1371/journal.pcbi.1002708.

[60] Trott, Oleg, and Arthur J. Olson. "AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading." *Journal of Computational Chemistry* 31, no. 2 (January 30, 2010): 455–61. doi:10.1002/jcc.21334.

[61] Wang, Renxiao, Luhua Lai, and Shaomeng Wang. "Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction." *Journal of Computer-Aided Molecular Design* 16, no. 1 (January 2002): 11–26. doi:10.1023/A:1016357811882.

Chapter 3 Construction of Coq6 homology models and stability screening through molecular dynamics

1. Introduction

Homology modeling consists of four main steps: template search, template-target alignment, 3D model construction, and model evaluation. In the previous chapter we introduced the methods of each step and briefly identified the challenges specific to Coq6. In this chapter we will describe the practical application of the homology modeling and molecular dynamics methods selected for Coq6 as well their results.

Coq6 is a particularly challenging target for homology modeling because of its low sequence identity to known templates. In this chapter we will describe an iterative process of developing Coq6 homology models to deal with the uncertainties arising from low sequence identity.

2. Template search

The general goal of homology modeling is to create a mapping between the residues of the target (the protein of unknown structure) and the residues of the template (the protein of known structure), and to use this sequence-to-sequence mapping to generate a 3D model of the target protein. As described in the previous chapter, the quality of the homology model depends on the similarity between the target and the template. When the two sequences have a sequence identity higher than 40%, 90% of backbone atoms can be modeled with an RMSD error of approximately 1Å with respect to the experimental structure.[1] When the sequence identities are less than 25%, there is only a 10% chance that the two sequences encode structurally homologous proteins.

2.1 Sequence based search: Phyre2[2]

Our initial Coq6 template search used the Phyre2 comparative modeling server[2], which uses a hidden Markov model (HMM) representation of the target protein sequence. This is a more generalized representation of the target protein's exact amino acid sequence, which enables HMM-based methods to recognize similarities between proteins with low sequence identities. The sequence profiles are used to generate pairwise alignments, and the search results are ranked according to this raw alignment score. The top 20 results of this search are presented below in **Table 3.1**.

TABLE 3.1: *Top 20 template hits specified with their PDB codes found by Phyre2's hidden Markov model (HMM) ranked by a combination of sequence coverage and sequence identity from the HMM alignment.*

Rank	PDB code	Coverage (%)	Sequence ID (%)	Function
1	2DKH	91	18	3-HB monooxygenase
2	2QA2	93	17	polyketide monooxygenase
3	4K2X	93	19	polyketide monooxygenase
4	2QA1	94	18	polyketide monooxygenase
5	3FMW	93	22	premithramycin B monooxygenase
6	1PN0	88	20	phenol 2- monooxygenase
7	2R0G	92	21	7-carboxy-K252c monooxygenase
8	3IHG	88	16	aklavinone-112 monooxygenase
9	4N9X	94	29	DDMQ6 monooxygenase
10	1PHH	93	15	p-hydroxybenzoate hydroxylase
11	4K22	87	28	Q C5 monooxygenase
12	4BK2	91	16	3-HB 6- monooxygenase
13	3E1T	94	12	chondrochloren halogenase
14	3I3L	88	14	Alkylhalidase
15	3NIX	93	15	Unknown
16	1K0I	93	23	p-hydroxybenzoate hydroxylase
17	2X3N	88	22	alkyl-quinolone monooxygenase
18	3ALL	87	21	MHPC monooxygenase
19	2RGJ	86	15	5-methyl PCA monooxygenase
20	3GMB	86	20	MHPCO monooxygenase

Despite having such low sequence identity, the templates found for Coq6 through sequence based searches are structurally very similar. They all have a PHBH-like global fold centered on the Rossmann fold for binding the FAD which faces a large beta sheet. This is shown graphically in **Figure 3.1** below.

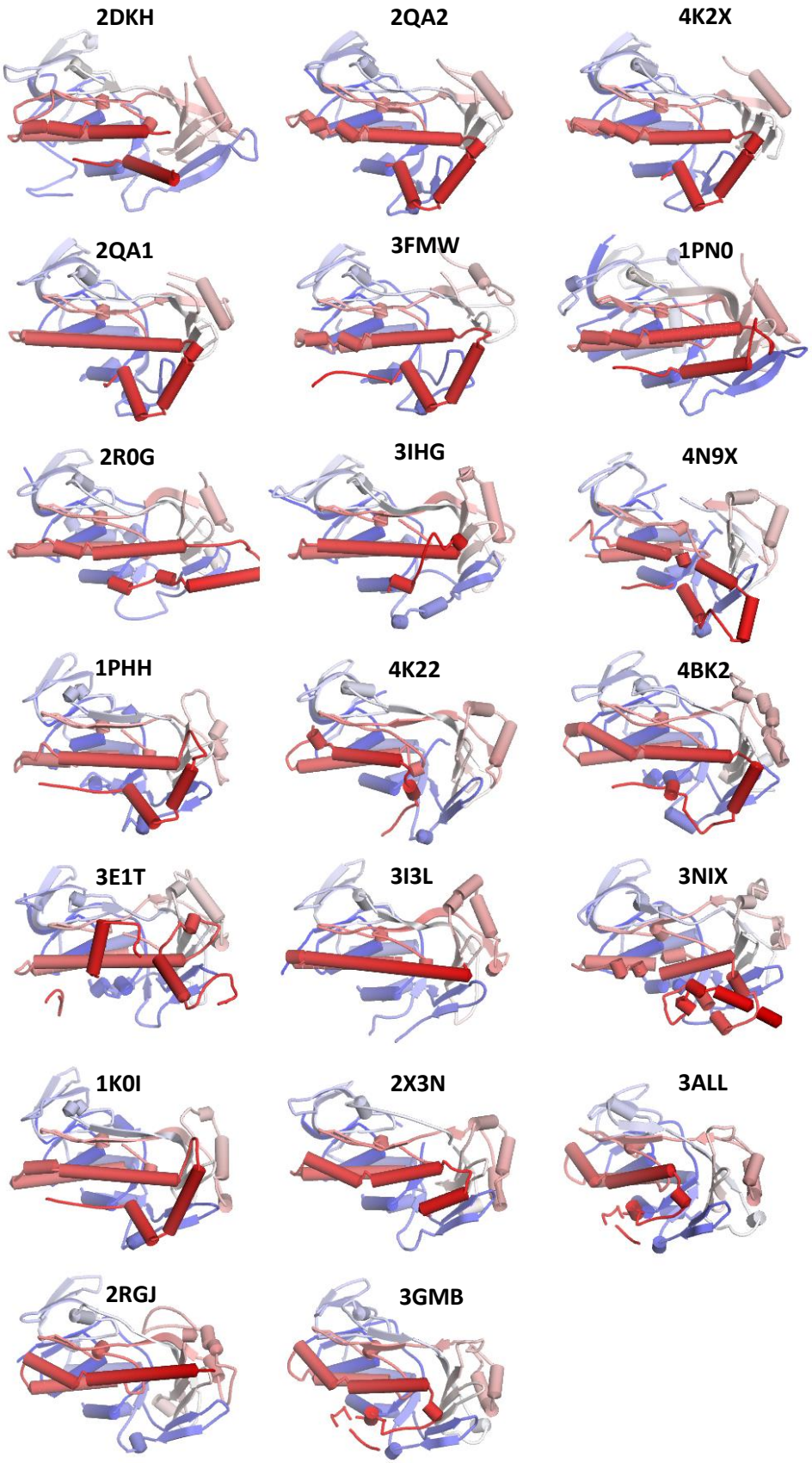

FIG 3.1: *3D structures of the top 20 template hits specified with their PDB codes found by Phyre2's hidden Markov model.*

The main reason for this is Coq6's requirement to bind the FAD cofactor in a specific conformation by forming specific contacts with it. An analysis of the FAD binding site in PHBH reveals that the majority of the enzyme-cofactor contacts involve the protein backbone.[3] This pattern of enzyme-cofactor contacts therefore imposes a strong constraint on the coordinates of the backbone if it is to result in a catalytically functional enzyme and is likely to be a major factor in conserving structure with such divergent sequences. However, the **low sequence identity immediately introduces four fundamental and inter-related challenges in selecting Coq6 templates**.

First, as protein structure is often more conserved than sequence,[4] the template with the highest global sequence identity score may not be the most structurally similar to the target. When all templates have identity scores that are similar with respect to each other and low with respect to the target, purely sequence based decision criteria are ambiguous. Second, given a set of low identity templates, the initial pairwise alignment between any given template and target may easily be sub-optimal. This means that the pairwise sequence identity computed between any template and the target may itself be erroneous. This leads directly to the third problem, which is that a list of templates ranked by this computed sequence identity may be in the wrong order. Finally, even if a single "correct" template ranking could be obtained, low *global* sequence identity may mean that different regions of the target's structure have diverged from the template by a different amount. That is to say, sequence identity, which is used to infer global structural similarity, may show significant variation over the extent of the target's sequence. This implies that certain templates may be better suited at modeling specific target regions, but not necessarily the entire target structure.

These challenges require a more careful analysis of the possible templates found through sequence based searching. **We must incorporate experimental knowledge into rational decisions for template selection.** At this point, our review of the potential templates on the basis of sequence analysis alone is likely to remain ambiguous. This is because **in the regime of low sequence identity (less than 30%), evolution has created more sequence variability than sequence conservation**, even *for the same global fold*. Yet this is a well-known and fundamental result of the study of protein structure: structure is more conserved than sequence.

This fundamental feature of protein structure works in our favor when we seek to align targets and templates that are already quite similar. This is because while sequence alignment algorithms generally operate by rewarding the matching of chemically similar residues**, *the structural implication of sequence alignment is that it is actually secondary structure elements that are being aligned***. That is to say, in the context of homology modeling, sequence alignment algorithms are really using chemical similarity of residues as a proxy for secondary structure alignment. When protein sequences are already similar there are therefore *two* elements working in our favor: similarity in sequence *and* similarity in secondary structure.

This fundamental feature of protein structure works against us when we seek to align targets and templates that are very different. This is because target and template may have diverged in two ways simultaneously, both in sequence *and* secondary structure. Therefore an alignment algorithm that seeks to maximize the matching of chemical similar residues may not accurately match up the respective secondary structure elements for the fundamental physical reason that these secondary structure elements *do not match*. Indeed, the matching of chemically similar residues can guide the alignment away from the "true" matching. Alternatively, an accumulation of complementary mutations between the target and template can produce a situation where secondary structure elements are structurally

well conserved but have diverged greatly in sequence. Again, the matching of chemically similar residues can guide the alignment away from a more structurally optimal matching.

In the twilight zone we must look beyond **alignment algorithms**. They perform well when evolutionary conservation of both sequence *and* structure have kept the target and template similar, leaving them **mainly the task of local optimization of chemical similarity**. However, they become **less useful when evolution has started to produce divergence of either sequence *or* structure**, and sometimes evolution produces **both simultaneously**. This is why we must also incorporate experimental knowledge into rational decisions for template selection.

Fortunately, late 2013 saw the publication of **two partial crystal structures of bacterial Q biosynthesis monooxygenases (deposited under PDB codes 4K22[5] and 4N9X[6]).** These are the closest functionally homologous proteins to Coq6, also a Q biosynthesis monooxygenase. This gives us a structural basis for template searching, which can give us better results, as well as additional information for interpreting the sequence based search results.

2.2 Structure based search: DALI[7]

Just as the results of a sequence based template search depend on the query sequence, the results of a structure based template search depend on the query structures. In our case, the query structures are two bacterial FAD dependent Q biosynthesis monooxygenases, PDB entries 4K22[5] and 4N9X[6]. We used the DALI server[7] to run two independent searches using either the 4K22 or 4N9X PDB structures as input. We will briefly present the top results of this search before describing each protein structure in more detail.

TABLE 3.2: *Top 5 template hits from DALI's structure-based search. Results are ranked by their RMSD Z-scores, which indicate the statistical significance of the computed RMSDs. Higher Z-scores indicate closer structural matches.*

DALI search using 4K22 as input					
Rank	Structure	Z-score	RMSD	Sequence ID	Function
1	4K22[5]	61.6	0	100	Q biosynthesis monooxygenase
2	4N9X[6]	38.9	1.8	67	Q biosynthesis monooxygenase
3	2X3N[8]	32	2.7	22	alkyl-quinolone monooxygenase
4	1BF3[9]	31.8	3	17	para-hydroxybenzoate hydroxylase
5	3E1T[10]	30.3	3.6	17	chondrocloren halogenase

DALI search using 4N9X as input					
Rank	Structure	Z-score	RMSD	Sequence ID	Function
1	4N9X[6]	64.5	0	100	Q biosynthesis monooxygenase
2	4K22[5]	38.9	1.8	67	Q biosynthesis monooxygenase
3	1DOE[11]	28.7	3.2	17	para-hydroxybenzoate hydroxylase
4	3NIX[12]	28.5	3.3	13	putative flavoprotein dehydrogenase
5	2X3N[8]	26.8	3	20	Alkyl-quinolone monooxygenase

We can see that 4K22 and 4N9X have the highest mutual geometric similarity, as they are each other's top result. However, both 4K22 and 4N9X have structural details (such as conformational distortions and missing coordinates) which affect their applicability as direct templates for homology modeling of Coq6. Therefore we will review the structure function features present in the top templates. The results of this review will inform our subsequent choices in templates for modeling Coq6. **We note the presence of para-hydroxybenzoate hydroxylase, a key reference enzyme of this family. It is represented in the DALI search by PDB structure 1BF3, which is a crystal structure of an R42K- C116S double mutant. We would prefer to use a structure of the wild-type PHBH if necessary in any modeling, represented by PDB structure 1PBE, which has the added value of being co-crystallized with its substrate.**

2.3 Top templates: a structural review

In this section we will review the main structural features of the top four templates identified through the structure based search. **Figure 3.2**, presented below, resumes these main structural features

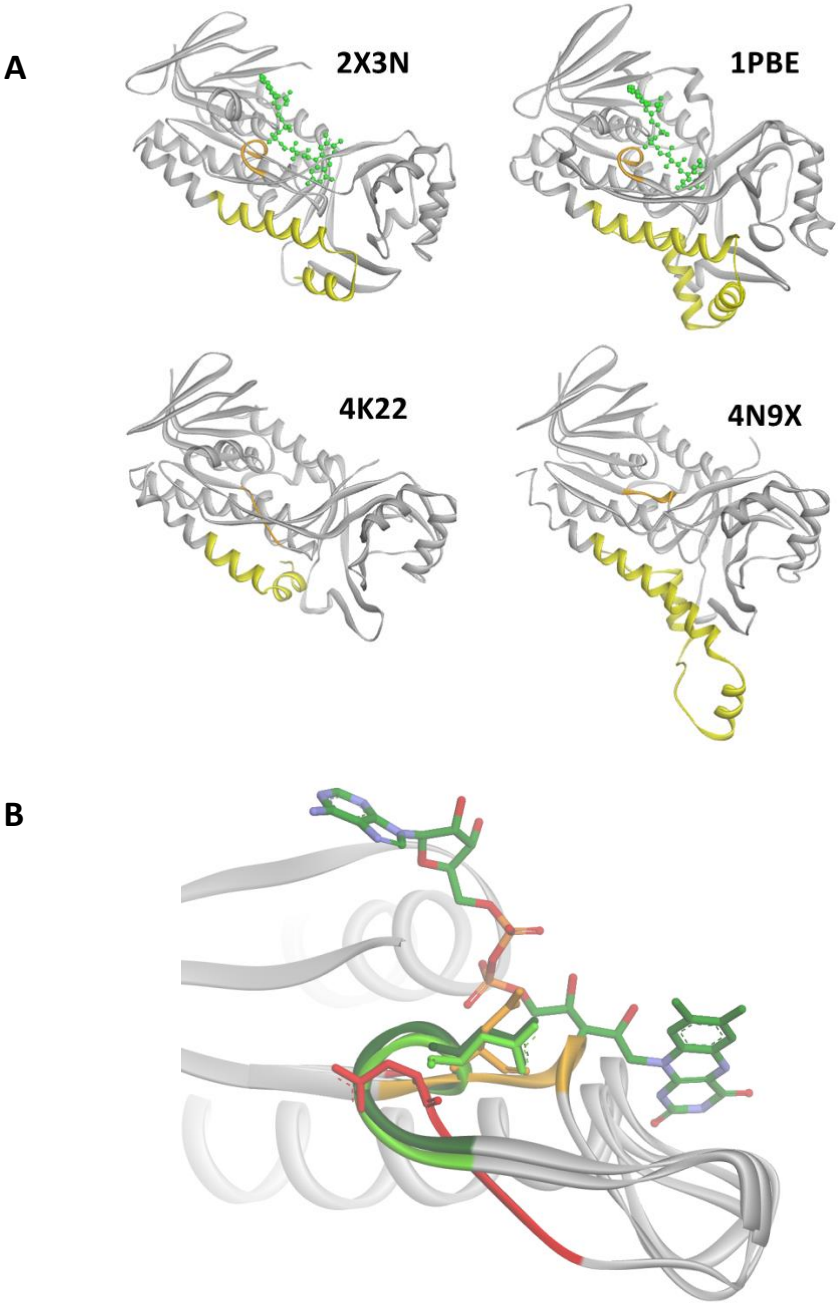

FIG 3.2: *A) Comparison of top template structures of Coq6. The N-terminal region is in grey, the C-terminal region is in yellow. The GDAxH motif is in orange. The FAD, when present in the experimental structure, is rendered in green sticks. B) Superposition of the GDAxH motif from the four templates. Dark green: 2X3N. light green: 1PBE, orange: 4N9X, red: 4K22. The motif forms a well-structured helix in 1PBE and 2X3N, which orients the motif's aspartate (shown as sticks) towards the FAD ribityl oxygens. In 4N9X,*

crystallized without FAD, the motif has a different conformation and would create a clash with an FAD assuming an extended conformation as observed in catalytically functional holo-enzymes such as PHBH. In 4K22, also crystallized without FAD, the motif is in yet a different conformation, orienting the aspartate away from its normal position.

2.3.1 4K22

28% sequence ID with Coq6

In late 2013 experimental researchers in our laboratory succeeded in crystallizing a proteolytically truncated and enzymatically inactive UbiI protein from *E. coli* (deposited under PDB code 4K22[5]). UbiI is the Q biosynthesis C-5 hydroxylase from *E. coli*, and functionally homologous to *S. cerevisiae* Coq6. The coordinates revealed that the construct was missing 35 C-terminal residues (some of which were likely to form part of the substrate binding region), as well as FAD. We infer that these features are responsible for the truncated UbiI being catalytically inactive.

Inspection of the top templates from Phyre2 indicates that the C-terminus of this class of enzyme (a Class A flavoprotein monooxygenase) is generally involved in closing the active site from below the plane of the Rossmann helices. The absence of the C-terminus in 4K22 leaves the active site open to bulk solvent, a feature known to be incompatible with catalysis from studies of PHBH.[13]

A broader comparison against other templates of the same global fold (member of the "PHBH-like" family as defined in SCOP,[14] the Structural Classification of Proteins, of the Class A flavoprotein monooxygenase,[15] and of the GR2[16]) reveals that 4K22 has an FAD binding site that is conformationally different from any template co-crystallized with FAD. In this class of FAD binding proteins the bottom of the FAD binding pocket is formed in part by a beta strand containing a GDAxH motif (represented in orange in **Figure 3.2**), which forms a single turn of alpha helix. This single turn of alpha helix orients the motif's aspartate side-chain towards the ribityl oxygens of FAD to form hydrogen bonds. This sequence motif always has the alpha helix conformation when its parent enzyme is co-crystallized with FAD. The deformation of this secondary structure element in experimentally solved members of this family, and more specifically in 4K22, seems to be incompatible with FAD binding, as GR2 enzymes of with a deformed GDAxH helix are never reported as co-crystallized with FAD in the PDB.

Together, the lack of FAD, a deformed FAD binding site, and an incomplete active site make this construct enzymatically inactive. It does not seem plausible to use the coordinates of a catalytically inactive enzyme to model Coq6. Therefore, we decided to exclude 4K22 as a direct source of coordinates for modeling Coq6. However, it is still a valuable starting point for a structure based template search.

2.3.2 4N9X

29% sequence ID with Coq6

Late 2013 also saw the deposition of another bacterial Q biosynthesis monooxygenase structure from *E. carotovora*, deposited under PDB code 4N9X.[6] The structure was solved as part of a Northeast Structural Genomics Consortium campaign. Despite the lack of an accompanying publication, the 4N9X structure is titled "Crystal structure of the **octaprenyl-methyl-methoxy-benzq** molecule from *Erwina carotovora* subsp. atroseptica strain SCRI 1043 / ATCC BAA-672, Northeast Structural Genomics Consortium (NESG) Target EwR161)", which implies a certain knowledge of the enzyme's substrate. The nomenclature chosen for this PDB entry tells us that the implied substrate is C-methylated at the C2 position and also

bears a methoxy group. However, it does not specify whether the methoxy group is on the C5 or C6 position, leaving two possibilities for this enzyme's substrate.

The 4N9X coordinates reveal that this construct is also missing FAD and has a deformed GDAxH loop. It is also missing coordinates for 15 residues comprising the *re* face edge beta strand due to excessive mobility. It is not clear if this is structurally related to the deformed GDAxH helix, or if it is an independent feature of this enzyme. However, this structure does bring us structural information for an important region missing from 4K22: the 35 C-terminal residues. While the coordinates of the N-terminal majority of 4N9X are not likely to be directly usable as a template for homology modeling, the coordinates of the C-terminus are valuable because they are missing from the closest structural and functional homologs, 4K22 and 2X3N.

The conformation of the 4N9X C-terminus is likely to be functionally relevant when we consider the evolutionary residue conservation of the C-terminus among members of the Coq6 family. This analysis (which is developed further in Chapter 4 – Selection of Coq6 models through molecular dynamics and substrate docking) identifies a stretch of residues as being highly conserved, suggesting a functional importance.

2.3.3 2X3N

22% sequence ID with Coq6

2X3N is the structure deposited for the *Pseudomonas aerguinosa* enzyme PqsL, a member of the quinolone signaling molecule synthesis pathway.[17] This makes it an especially interesting template because of its likely function as an alkyl-quinolone monooxygenase, which means it has evolved to hydroxylate a substrate with structural similarity to Q biosynthesis intermediates. The 2X3N structure also features a well formed GDAxH helix and is bound to FAD. The only drawback to using 2X3N as a single template for modeling Coq6 is the lack of coordinates for the C-terminus of the enzyme, (residues 371-398) which is generally predicted to form part of the active site. Fortunately, the 4N9X structure presented above does have resolved coordinates for the C-terminus, giving us a template for this region from an even closer functional homolog.

2.3.4 1PBE (PHBH, or para-hydroxybenzoate hydroxylase)

15% sequence ID with Coq6

The Phyre2 search return PHBH (structure 1PHH[18]) as one of the possible templates. Despite its low sequence identity, we include it in this section because all of its functional residues have been resolved, it has been co-crystallized with its substrate, and there is extensive biochemical characterization of it in the literature. However, structure 1PHH is a mutant form of PHBH complexed with its product. For the purposes of homology modeling we want to use a structure of the wild-type PHBH co-crystallized with its substrate, since this configuration is closer to the type of homology model we intend to produce for Coq6. 1PBE[19] is a crystal structure of the wild-type PHBH complexed with FAD and its substrate. The general structure-function description for this extensively characterized enzyme[20] has already been given in Chapter 1. 1PBE has a lower structure similarity Z-score to 4K22 (31.7) and 4N9X (38.9) than 2X3N (32), and its substrate does not have an aliphatic chain, as do Q biosynthesis intermediates, or the putative PqsL substrate (an alkylated quinolone), and so therefore it is not a first choice as a template for modeling Coq6. This is particularly true with regard to the conformation of the C-terminus, which differs from the conformation seen in 4N9X, a closer functional relative. However, the substrate, para-hydroxybenzoate, is a partially-substituted six-membered aromatic ring, which is similar to the quinone

moiety of Q biosynthesis intermediates. Given that these conformations of PHBH are enzymatically active, this crystal structure shows us an important example of what a catalytically plausible substrate position looks like. In particular, we can use the geometry of the enzyme-FAD-substrate complex to inform the construction and interpretation of substrate docking in our Coq6 homology model (as described in Chapter 4 – Selection of models by substrate docking).

2.4 The Coq6 global fold can be divided into two regions for homology modeling: N-terminus and C-terminus

The important conclusion of this analysis is that the global fold of Coq6 may be divided into two regions for the purposes of homology model construction:

- an N-terminal majority of the protein which contains the FAD binding domain. The key criterion for choosing an N-terminal template is the ability to bind FAD, as evidenced by the presence of FAD in the crystal structure, and the proper formation of the associated GDAxH helix.

- a smaller C-terminal region which is likely to form part of a substrate binding site. The criterion for choosing a C-terminal template is not immediately clear, as we have no examples of homologous enzymes co-crystallized with substrates homologous to Q biosynthesis intermediates. However, it must satisfy the more generic criterion of generating a structurally stable conformation. It must also contribute to the formation of a stable substrate binding region allowing the aromatic center of Q biosynthesis intermediates to reach a catalytically plausible position in front the FAD isoalloxazine.

Structural stability is addressed in the second half of this chapter through molecular dynamics simulations. Substrate binding ability is addressed in Chapter 4 (Selection of Coq6 models through molecular dynamics and substrate docking).

From our review it therefore seems clear that only 2X3N and 1PBE could serve as functional templates for the N-terminus of Coq6. Between these two choices, the higher structural similarity of 2X3N (as calculated by the DALI method) to known Q biosynthesis hydroxylases would indicate that it is the best template for modeling the N-terminal majority of Coq6. Despite this conclusion, we will still test 4K22 as a template for the N-terminus of the protein in our larger panel of homology models in order to explore any potential valorization of this enzymatically inactive structure. Our review also indicates that among these templates, only two can be considered for modeling the C-terminus: 1PBE and 4N9X, since 2X3N and 1PBE are missing coordinates for their C-termini.

However, before building homology models based on these template combinations, we must also consider a large region of the Coq6 sequence with no structural precedent that is difficult to align with any of these templates. The sequences of Coq6 and the other top templates are presented below. The top templates have sequences that are about 400 residues long, whereas the Coq6 sequence is 479 residues long. Even subtracting the first 17 residues as part of the mitochondrial signal sequence the Coq6 sequence is still about 50 residues longer than its best templates. At this stage, we cannot reliably assign residues in an alignment, although we can detect three conserved sequence motifs common to PHBH-like proteins[21] in **Figure 3.3** below: the ADP binding motif (GxGxxG) highlighted in orange, the NAD(P)H binding motif (GxDGxxx) highlighted in blue, and the ribityl binding motif (GDAxH) highlighted in green. While we have not yet performed an explicit sequence alignment, it is visually apparent that the region between the NAD(P)H binding motif and the ribityl binding motif is significantly larger in Coq6

than in any of its templates. This region is highlighted in purple in **Figure 3.3** below. In Coq6 this region contains 167 residues, whereas in the templates from *E. carotovora* and *E. coli* it is 119 residues; in PHBH it is 121 residues, and in the alkylquinolone hydroxylase from *P. aerguinosa* it is 125 residues.

Coq6 (479 residues)

MFFSKVMLTRRILVRGLATAKSSAPKLTDVLIVGGGPAGLTLAASIKNSPQLKDLKTTLV
DMVDLKDKLSDFYNSPPDYFTNRIVSVTPRSIHFLENNAGATLMHDRIQSYDGLYVTDGC
SKATLDLARDSMLCMIEIINIQASLYNRISQYDSKKDSIDIIDNTKVVNIKHSDPNDPLS
WPLVTLSNGEVYKTRLLVGADGFNSPTRRFSQIPSRGWMYNAYGVVASMKLEYPPFKLRG
WQRFLPTGPIAHLPMPENNATLVWSSSERLSRLLLSLPPESFTALINAAFVLEDADMNYY
YRTLEDGSMDTDKLIEDIKFRTEEIYATLKDESDIDEIYPPRVVSIIDKTRARFPLKLTH
ADRYCTDRVALVGDAAHTTHPLAGQGLNMGQTDVHGLVYALEKAMERGLDIGSSLSLEPF
WAERYPSNNVLLGMADKLFKLYHTNFPPVVALRTFGLNLTNKIGPVKNMIIDTLGGNEK

E. carotovora Q hydroxylase (403 residues, full sequence of PDB entry 4N9X)

MQSFDVVIAGGGMVGLALACGLQGSGLRIAVLEKQAAEPQTLGKGHALRVSAINAASECL
LRHIGVWENLVAQRVSPYNDMQVWDKDSFGKISFSGEEFGFSHLGHIIENPVIQQVLWQR
ASQLSDITLLSPTSLKQVAWGENEAFITLQDDSMLTARLVVGADGAHSWLRQHADIPLTF
WDYGHHALVANIRTEHPHQSVARQAFHGDGILAFLPLDDPHLCSIVWSLSPEQALVMQSL
PVEEFNRQVAMAFDMRLGLCELESERQTFPLMGRYARSFAAHRLVLVGDAAHTIHPLAGQ
GVNLGFMDVAELIAELKRLQTQGKDIGQHLYLRRYERRRKHSAAVMLASMQGFRELFDGD
NPAKKLLRDVGLVLADKLPGIKPTLVRQAMGLHDLPDWLSAGK

E. coli Q hydroxylase (400 residues, full sequence of PDB entry 4K22)

MQSVDVAIVGGGMVGLAVACGLQGSGLRVAVLEQRVQEPLAANAPPQLRVSAINAASEKL
LTRLGVWQDILSRRASCYHGMEVWDKDSFGHISFDDQSMGYSHLGHIVENSVIHYALWNK
AHQSSDITLLAPAELQQVAWGENETFLTLKDGSMLTARLVIGADGANSWLRNKADIPLTF
WDYQHHALVATIRTEEPHDAVARQVFHGEGILAFLPLSDPHLCSIVWSLSPEEAQRMQQA
SEDEFNRALNIAFDNRLGLCKVESARQVFPLTGRYARQFASHRLALVGDAAHTIHPLAGQ
GVNLGFMDAAELIAELKRLHRQGKDIGQYIYLRRYERSRKHSAALMLAGMQGFRDLFSGT
NPAKKLLRDIGLKLADTLPGVKPQLIRQAMGLNDLPEWLR

P. fluorescens pHB hydroxylase (394 residues, full sequence of PDB entry 1PBE)

MKTQVAIIGAGPSGLLLGQLLHKAGIDNVILERQTPDYVLGRIRAGVLEQGMVDLLREAG
VDRRMARDGLVHEGVEIAFAGQRRRIDLKRLSGGKTVTVYGQTEVTRDLMEAREACGATT
VYQAAEVRLHDLQGERPYVTFERDGERLRLDCDYIAGCDGFHGISRQSIPAERLKVFERV
YPFGWLGLLADTPPVSHELIYANHPRGFALCSQRSATRSRYYVQVPLTEKVEDWSDERFW
TELKARLPAEVAEKLVTGPSLEKSIAPLRSFVVEPMQHGRLFLAGDAAHIVPPTGAKGLN
LAASDVSTLYRLLLKAYREGRGELLERYSAICLRRIWKAERFSWWMTSVLHRFPDTDAFS
QRIQQTELEYYLGSEAGLATIAENYVGLPYEEIE

P. aerguinosa alkykquinolone hydroxylase (full sequence of PDB entry 2X3N)

MTDNHIDVLINGCGIGGAMLAYLLGRQGHRVVVVEQARRERAINGADLLKPAGIRVVEAA
GLLAEVTRRGGRVRHELEVYHDGELLRYFNYSSVDARGYFILMPCESLRRLVLEKIDGEA
TVEMLFETRIEAVQRDERHAIDQVRLNDGRVLRPRVVVGADGIASYVRRRLLDIDVERRP
YPSPMLVGTFALAPCVAERNRLYVDSQGGLAYFYPIGFDRARLVVSFPREEARELMADTR
GESLRRRLQRFVGDESAEAIAAVTGTSRFKGIPIGYLNLDRYWADNVAMLGDAIHNVHPI
TGQGMNLAIEDASALADALDLALRDACALEDALAGYQAERFPVNQAIVSYGHALATSLED
RQRFAGVFDTALQGSSRTPEALGGERSYQPVRSPAPLG

FIG 3.3: *The top templates, as well as Coq6, all come from the PHBH-like family. The family contains three easily recognizable sequence features. In orange, the GxGxxG motif. In blue, the GxDGxxx motif. In green, the GDAxH motif. The top templates are about 400 residues long, whereas the Coq6 sequence is 479 residues. The segment between the GxDGxxx motif and the GDAxH, highlighted in purple, is the region likely to contain the ~50 extra residues of Coq6. The Coq6 mitochondrial signal sequence is highlighted in beige.*

2.5 Coq6 contains an additional subdomain not present in known structural homologs

Comparison of the sequences of Coq6 and the various bacterial homologs immediately reveals that the Coq6 sequence in *S. cerevisiae* is significantly longer: it contains about 80 extra residues. The N-terminus of the sequence contains the mitochondrial localization signal and consists of 17 residues (MFFSKVMLTRRILVRGL), a region absent in the bacterial homolog templates. This leaves about 50 extra residues which have no homologous residues among the bacterial templates. The complete primary sequence of 4K22 (UbiI) contains 400 residues; 4N9X (a putative Q biosynthesis intermediate) contains 403 residues; 2X3N (a putative alkyl-quinolone hydroxylase) contains 398; 1PBE (p-hydroxybenzoate hydroxylase) contains 394.

Sequence landmarks for three highly conserved sequence motifs (as shown in **Figure 3.3**) allow us to roughly estimate the positioning of these 50 extra residues as occurring between the GxDGxxx and GDAxH motifs. Preliminary alignments made between Coq6 and 4K22 suggested that these additional residues formed a single continuous block called the Coq6-family insert, and was likely to form a structurally peripheral element of the protein structure, since such a region is not observed in the global fold of any Coq6 structural templates.

However, the target-template pairs are too dissimilar to unambiguously produce an optimal alignment for each. Comparing Coq6 to only the bacterial templates cannot help us accurately delimit the edges of the insert. Fortunately, comparison of *S. cerevisiae* Coq6 to other Coq6 sequences in a multiple sequence alignment (MSA) can help us make this distinction more clearly, as detailed below.

2.6 A Coq6-family MSA helps define the insert sequence

There is a strong physical basis for using an MSA of the Coq6 family to help define the insert. The "minimal" global fold of this class of enzymes (as determined from experimental structures in the PDB, typically bacterial examples such as 1PBE) does not require any additional subdomain (such as might be encoded by the Coq6 insert) to be catalytically functional. Therefore, we posit that even within the Coq6 family of proteins, this insert region is likely to display greater structural variability than the rest of the sequence. By comparing the sequence of yeast Coq6 to other Coq6 proteins in a Coq6-family multiple sequence alignment (see **Figure 3.4**), we are likely to gain clues about structural position of the insert. This Coq6-family MSA will then help us to produce better target-template alignments. The complete MSA was computed with ConSurf[22] and is presented in Annex 2. An excerpt is presented in **Figure 3.4** here to show how the MSA allowed us to identify the Coq6 insert in a rather clear way.

N-terminus

Coq6 family insert

C-terminus

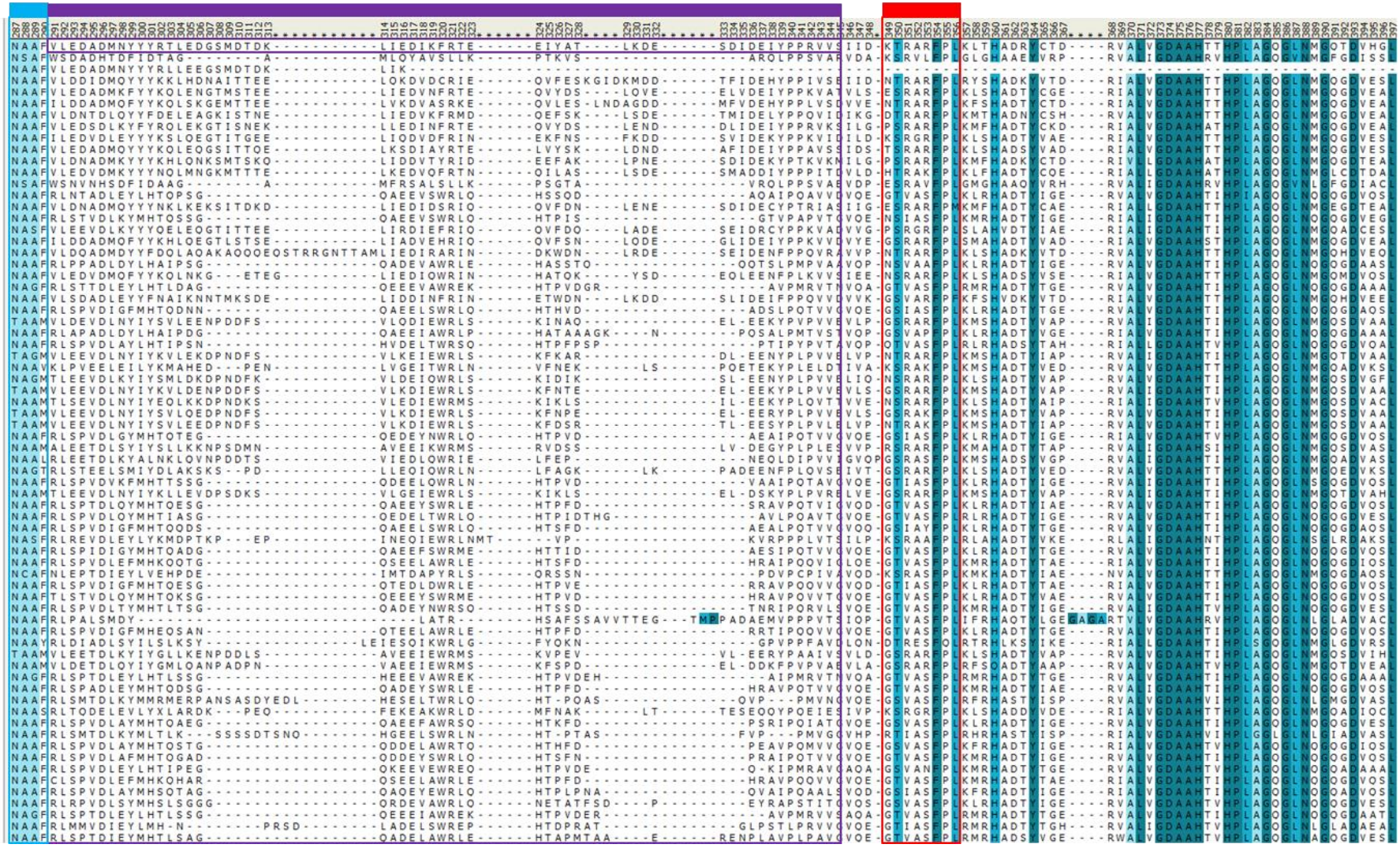

FIG 3.4: *(Preceding page) Excerpt of the Coq6 MSA. The S. cerevisiae Coq6 sequence is the top line. The H. sapiens Coq6 sequence is the second line from the top. The MSA allows us to recognize the insert as being flanked by two conserved sequence motifs: an N-terminal NAA motif in the light blue frame, and a C-terminal SxASFPL motif in the red frame. Sequence numbering is for the S. cerevisiae Coq6 sequence.*

It is apparent that the insert is of very variable length and composition, but that its N-terminal and C-terminal borders contain sequence motifs that show some key conserved residues. It is residue conservation at these points that helps us determine the likely borders, and therefore the size, of the insert in the *S. cerevisiae* Coq6 sequence.

At this point of our review of the template structures with respect to the Coq6 sequence we have two main goals. The first is to test the stability of several multi-template models based on different template combinations without the insert. Once we have identified the most stable combination of templates for modeling the core of Coq6, we will then create a set of models including the insert in order to test our hypotheses about its structure.

3. Model building

3.1 Modeling strategy: Construction of a combinatorial set of multiple template models

Our review of the templates and the Coq6 sequence itself indicates that there is likely not a single best template for creating a homology model. Rather, we can use specific regions from each template to model specific regions of Coq6, creating a model based on multiple-templates. These regions have been identified as the N-terminal majority of the protein (which houses the FAD binding site), and the C-terminus (which is likely to form part of the substrate binding site).

This segmented attribution of Coq6 residues to template coordinates is necessary for two reasons: some enzymatically functional templates have missing coordinates, and some templates are enzymatically inactive. In addition, the target contains the Coq6-family insert region, for which there is no template at all. While experimental knowledge of each template can guide us to a rational selection of specific template regions for modeling specific target regions, there remains the possibility that our selection may be too narrow. **Therefore, we will create a larger set of homology models to test various structural possibilities for each region of structure: the N-terminus, the C-terminus, and the insert**.

The iterative and exploratory nature of this phase of homology modeling building was defined by the development of three generations of homology models. The first generation was based on an initial sequence alignment of Coq6 against 4K22 and deliberately excluded the insert. Analysis of these models revealed an additional structural incompatibility between the Coq6 sequence and the 4K22 structure in the insert region. This was addressed by the creation of a second generation of homology models based on a sequence alignment of Coq6 against 2X3N, also excluding the insert. Finally, a third generation of models was created using second generation structures while adding the insert. The coordinate sources of the complete panel of homology models is presented below and divided into three generations.

TABLE 3.3: *Table of the multi-template homology model coordinate sources for the N-terminal and C-terminal regions of Coq6. Three generations of target-template alignments were made. The first generation alignments were based on Coq6 and 4K22; the second generation was based on Coq6 and 2X3N. Generation 1 models are not physically relevant and their dynamics are not presented here. Generation 2 models without the Coq6 insert were used to determine which template combinations gave stable models. Generation 3 models add the Coq6-family insert.*

Coq6 Target sequence	N-ter	C-ter
Generation 1 Without Insertion	4K22	none
	4K22	4N9X
	4K22	1PBE
Generation 2 Without Insertion	2X3N	1PBE
	2X3N	4N9X
Generation 3 Full Length	2X3N	4N9X
	2X3N	I-TASSER
	1PBE	ROBETTA

3.1.1 Generation 1: 4K22 as the Coq6 N-terminal template; no FAD or Coq6-family insert

Structural incompatibility between the 4K22 structure and the Coq6 sequence helps define the insert and select an optimal N-terminal template

The first generation of alignments was made between Coq6 and 4K22. This alignment suggested that the Coq6-insert was a single continuous block of residues.

FIG 3.5: *Identification of the Coq6 insert sequence by the initial Coq6-4K22 target-template alignment. This target-template alignment became obsolete upon further structural analysis of the resulting 4K22 based models.*

However, analysis of the resulting 4K22-based homology models revealed an incompatibility between the alignment and the resulting structure in the region of the insert. Although we do not yet know what the secondary or tertiary structure of the insert may be, we know it must satisfy one general geometric constraint: its structure must be compatible with the rest of the global fold. The region where this constraint is likely to be the strongest are the border regions where the insert sequence is predicted to rejoin the more minimal global fold observed in the templates. The Coq6-insert border regions are composed of the INAFF sequence on the N-terminal side and of the SIIDKT sequence on the C-terminal side. Therefore, we will examine the structure of the regions predicted to flank the insert on the experimentally solved templates in order to establish their native geometry (which depends on the

template's amino acid sequence), and then assess their ability to accommodate the geometry possible for the insert (which depends on the Coq6-insert amino acid sequence).

The region of structure predicted to accommodate the insert is the exterior of the beta sheet domain, as shown in **Figure 3.6** below. The exterior of the beta sheet is partially covered by another layer of structure composed of three general elements. Listed in the N-terminal to C-terminal progression, there is a descending motif, a turn motif, and an ascending motif. In both the 4K22 and 2X3N structures, the descending motif is a short stretch of helix, so this is not the discriminating structural factor for compatibility with the Coq6 sequence. This is followed by the turn motif, which serves to turn the backbone about 90 degrees to the descending helix axis and reach across the exterior of the beta sheet to rejoin the ascending motif. In 4K22, the ascending motif is a 4 residue beta strand, whereas in 2X3N it is a 9 residue alpha helix. Analysis of this region reveals that the turn motif is the discriminating factor in template structural compatibility with the Coq6 sequence.

In 4K22, the turn motif is composed of the DNRLG sequence whose backbone angles fall into sterically allowed regions (including the relatively rare left handed helix geometry) with the exception of the terminal G258, which is strictly conserved among the 150 closest UbiI homologs as computed by ConSurf. Therefore, we conclude that the turn motif geometry observed in 4K22 requires a G in this structural (and sequence alignment) position.

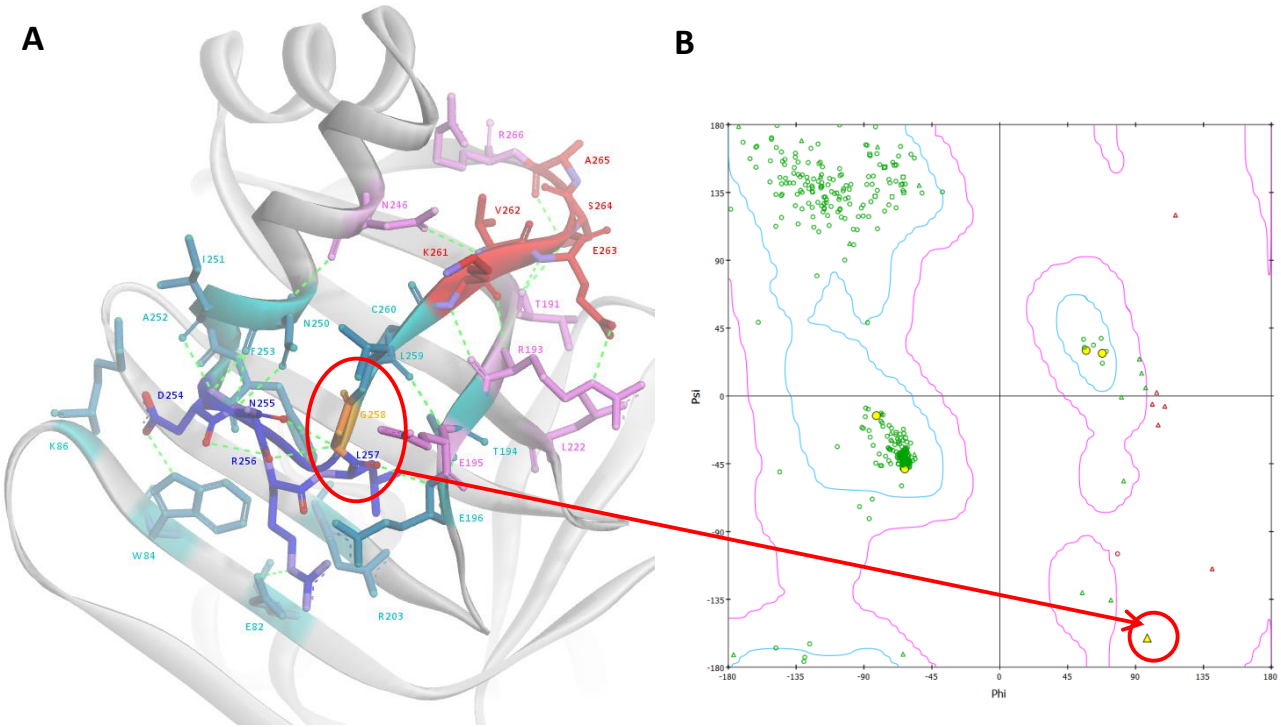

FIG 3.6 *The DNRLG motif turn geometry observed in 4K22 relies on a backbone angle forbidden to all residues except glycine; therefore a conserved G would be required in Coq6, but it is not present. A) UbiI's turn motif is in dark blue; the ascending motif is in red; G258 is highlighted in orange and B) on the Ramachandran plot.*

However, the Coq6 sequence does not contain a G at this position in the Generation 1 alignment. In fact, the Coq6 insert sequence contains only a single G, and it is at best 9 residues away from being aligned to a homologous position according to the 4K22 based Generation 1 alignment. Since no residue other than glycine could form this structure, this is a strong indication that this local region of 4K22 is not compatible with the Coq6 sequence. **This detail is another disqualifying factor (in addition to the lack of FAD, deformed GDAxH loop, and missing C-terminus) for the use of 4K22 as a direct modeling template.** Therefore, we will not present the first generation of 3D models, as they were not useful for further modeling. Instead, we will use the conclusions drawn from this analysis to create the second generation of target-template alignments used in creating homology models.

The conclusion drawn from this analysis is that using 4K22 as an N-terminal template will not generate a physically plausible model of the Coq6 sequence because of mal-formation of the GDAxH helix as well as the incompatibility of the Coq6 sequence with the 4K22 geometry in the region of the Coq6 insert.

3.1.2 Generation 2: 2X3N as the Coq6 N-terminal template; no FAD or Coq6-family insert

Structural compatibility between the 2X3N structure and the Coq6 sequence refines the insert alignment and gives clues to secondary structure

In 2X3N, the turn motif is composed of the VGDES sequence (as shown in **Figure 3.7**) whose backbone angles fall into the more common regions of the Ramachandran plot describing right handed alpha helices and beta strands. The 2X3N turn motif does not require any special residue (glycine in this case) to form its geometry. In addition, the ascending motif immediately following is longer and forms a short alpha helix. The slightly increased length of the ascending motif is an additional commonality with the Coq6, and it is interesting to see these extra residues organized into an alpha helix. Using ClustalO[23] and the Coq6 MSA (see **Figure 3.4**) to refine the alignment between Coq6 and 2X3N in the insert region, we arrive at a different conclusion: in the structural context of 2X3N, the Coq6 insert sequence is not a single continuous block with respect to this template. Instead, the alignment indicates that the Coq6 insert contains a region structurally homologous to the ascending element of 2X3N: an alpha helix consisting of the FRTEEIYAT sequence (see **Figure 3.7**).

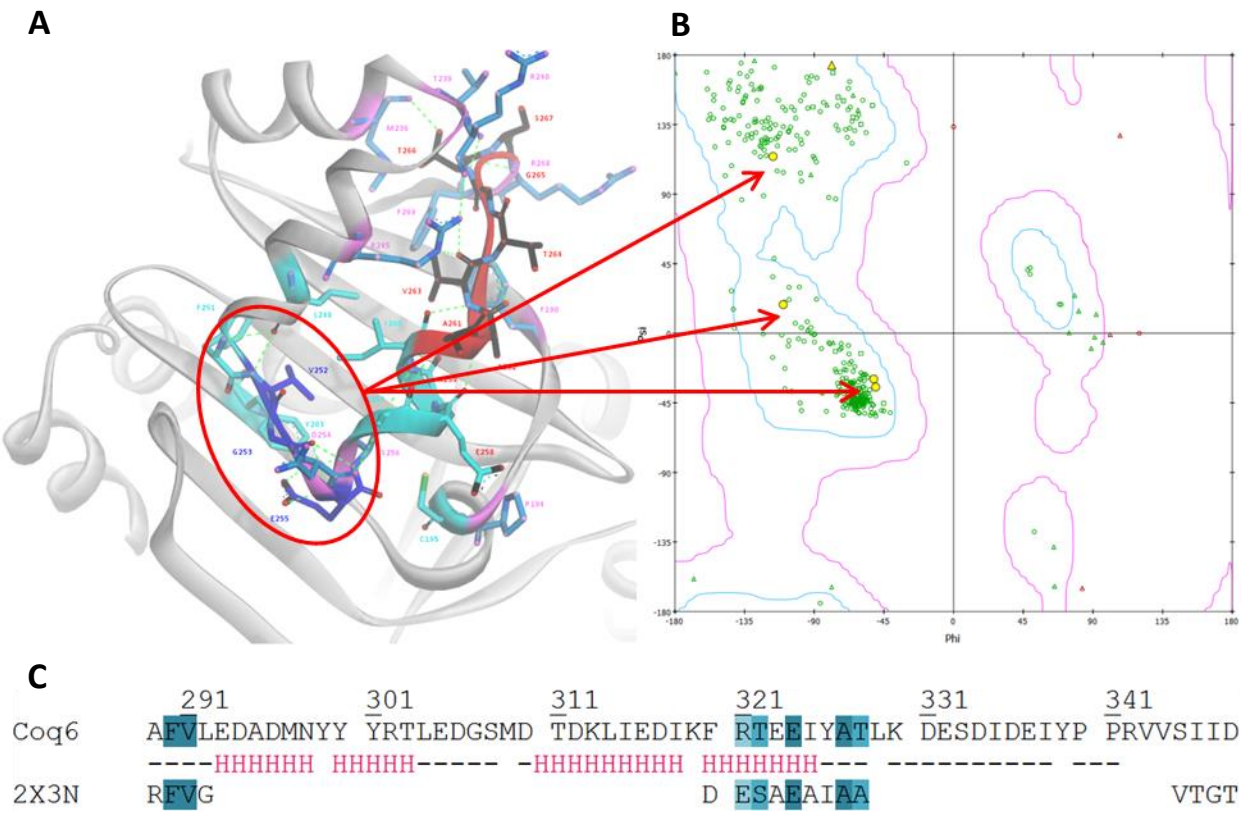

FIG 3.7 *VGDES turn geometry does not rely on any unusual backbone angles requiring glycine to be conserved. A) 2X3N's turn motif is in dark blue and the ascending motif is in red. B) Ramachandran plot showing the turn motif residues as yellow circles, indicated by arrows. C) The alignment in this region between Coq6 an 2X3N, including a secondary structure prediction for the Coq6 insert sequence.*

Additionally, the alignment of the 2X3N FRTEEIYAT sequence towards the middle of the Coq6 insert suggests that the additional residues comprising the insert may have evolved as N- and C-terminal elaborations of an existing helical ascending motif present in a more "minimal" bacterial ancestor enzyme, of which 2X3N may be an example.

The compatibility of the Coq6 insert sequence with the ascending helix motif of 2X3N is also supported by predictions of its secondary structure using the Jpred server.[24] This prediction assigns a secondary structure state (helix, sheet or coil) to each residue in the sequence submitted. This predicts a general helix-turn-helix motif, with the turn centered around the DG motif, which contains the only glycine in the insert sequence. These secondary structure assignments can be specified in the construction of homology models, in our case the MODELLER software. Jpred was also used to predict the secondary structure of the *si* loop as being mainly alpha-helical. This is consistent with experimental structures of this global fold where the *si*-loop is resolved.

Of the Generation 2 Coq6 models tested we conclude that only the 2X3N-4N9X based model presents a structure stable over 20ns of molecular dynamics. Fortunately, the 2X3N based moiety is also compatible with FAD binding for catalysis, since the original 2X3N crystal structure contains it. Therefore, we select the 2X3N and 4N9X structures as templates for generating models of Coq6. The next phase will be to integrate the Coq6-family insert to the enzyme model. Modeling of the Coq6 sequence without the insert is presented in Annex 5.

3.2 Homology models including the insert are used to design constructs for *in vivo* testing

The Coq6 insert is of particular interest because it is not present in enzymatically active bacterial homologs, and therefore would seem to have a function other than catalysis. Experimental knowledge of the CoQ synthome as an obligate multi-protein complex with functionally interdependent proteins[25] strongly suggests that the ability to form specific protein-protein interactions will be an important feature of Coq proteins. Our experimental partner in Grenoble, Dr. Fabien Pierrel, developed the working hypothesis that the function of the Coq6 insert was to mediate protein-protein interaction within the CoQ synthome. Dr. Pierrel planned a series of experiments to see if swapping the insert-regions of *S. cerevisiae* Coq6 with *H. sapiens* Coq6 would enable inter-species complementation of Coq6-null mutants. Therefore, a precise definition of the insert was needed in order to excise it without disrupting the structure or function of the enzyme. While the initial sequence definitions of the insert were based on the Coq6-MSA guided alignment, they needed to be subsequently refined by structural analysis of Coq6 homology models before producing insert definitions suitable for experimental testing. However, none of the proposed constructs showed activity *in vivo,* revealing its importance in order to maintain the protein's integrity. At this point we will therefore present the Generation 3 homology models with the insert, which will be used for further analysis in the rest of the thesis.

The next phase is the construction of full length models of Coq6 that include the insert. Our manually curated construction method will recapitulate the procedure used for the Generation 2 2X3N-4N9X based model, with the inclusion of the Coq6-family insert in the target sequence. We also included a short region of the 4K22 structure for one of the peripheral strands of the sheet-domain sheet in order to allow the FAD isoalloxazine enough room to undergo its in-out conformational movements. In addition, we also used two automated modeling servers for modeling the full length Coq6 sequence. First we will present the alignments used to construct the Generation 3 models. Then we will compare the resulting homology models before comparing their behavior in molecular dynamics simulations.

3.2.1 Generation 3: 2X3N as the Coq6 N-terminal template; with FAD and Coq6-family insert

In order to create a Coq6 model representative of a functional enzyme we decided to use different regions of each template to model different regions of Coq6. The manually curated multiple template alignment used to create the Generation 3 Coq6 model is presented below. This model is hereafter referred to as the **RATIONAL** Coq6 model.

FIG 3.8 *Sequence alignment of Coq6 with templates used in the Generation 3 manually curated alignment homology model. The templates are referred to by their PDB codes. The portions of each template used for constructing the Generation 3 Coq6 model are framed in green. Secondary structure as calculated by DSSP is indicated by text color: beta strands are blue, alpha helices are red, turns are in black. Residues missing from crystal structures have background highlighted in red. FAD binding residues are highlighted in green.*

3.2.2 Generation 3: Homology models from I-TASSER and ROBETTA

In order to investigate alternative methods for template selection, alignment, and model generation, we used the I-TASSER[26] and ROBETTA[27] automated servers using their default parameters. The I-TASSER and ROBETTA comparative modeling servers follow the same basic steps of template search, alignment, and model building. Both methods use sequence profile scoring methods to find and select the most appropriate templates. I-TASSER selected 2X3N as the template for Coq6; however, since 2X3N is missing coordinates for the C-terminus, I-TASSER reconstructed this region as well as the insert according to its own protocol.[26] ROBETTA selected 1PBE as its template for Coq6. Because 1PBE was crystallized with its C-terminus, the only major region constructed according to its specific protocol is the insert, whose secondary structure is predicted by DSSP.[28] The alignments used by each method are presented below in **Figures 3.9 and 3.10**.

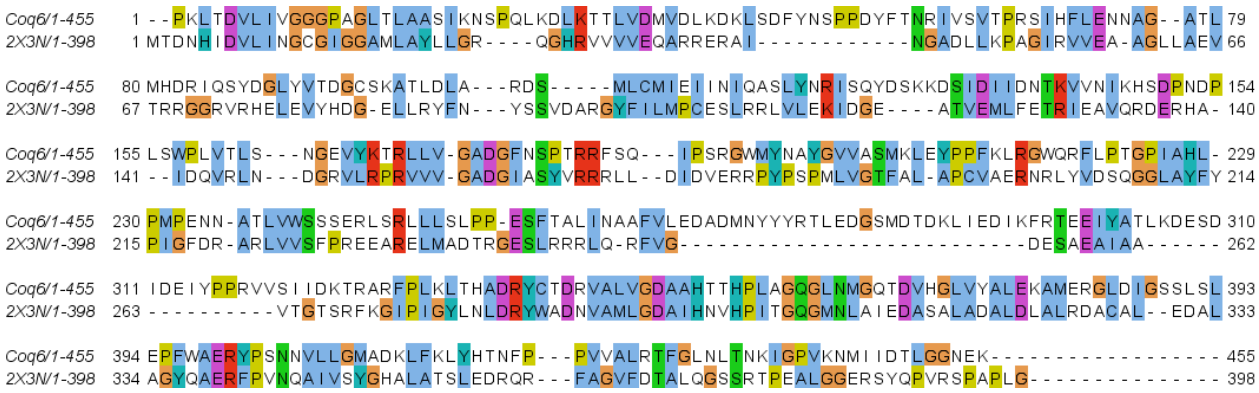

FIG 3.9 *The alignment used by I-TASSER to produce the Coq6-I-TASSER model, presented in the ClustalX color scheme, wherein residue G,P,S, and T are colored orange; H,K, and R are colored red; F,W, and Y are colored blue; I,L,M, and V are colored green.*

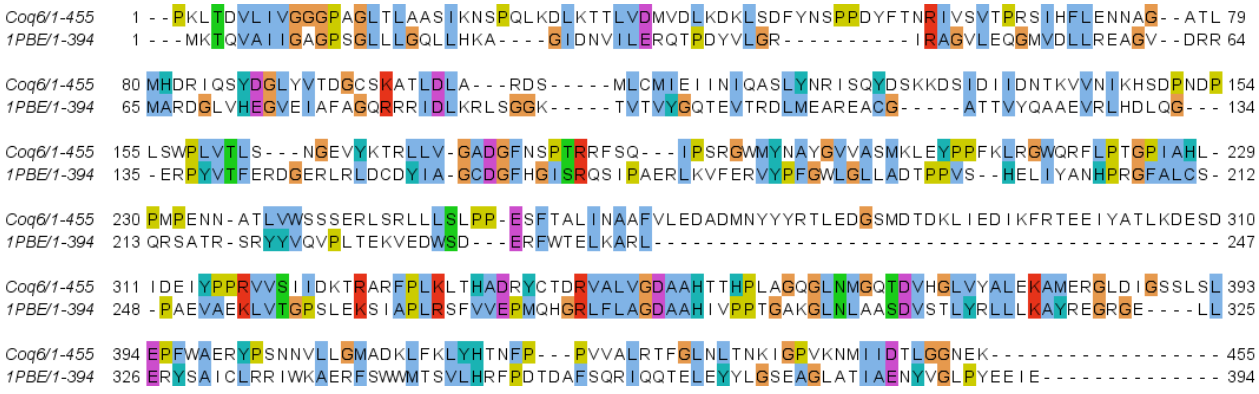

FIG 3.10 *The alignment used by ROBETTA to produce the Coq6-ROBETTA model, presented in the ClustalX color scheme, wherein residue G,P,S, and T are colored orange; H,K, and R are colored red; F,W, and Y are colored blue; I,L,M, and V are colored green.*

The 3D homology models resulting from each respective build procedure is presented in **Figure 3.11** below.

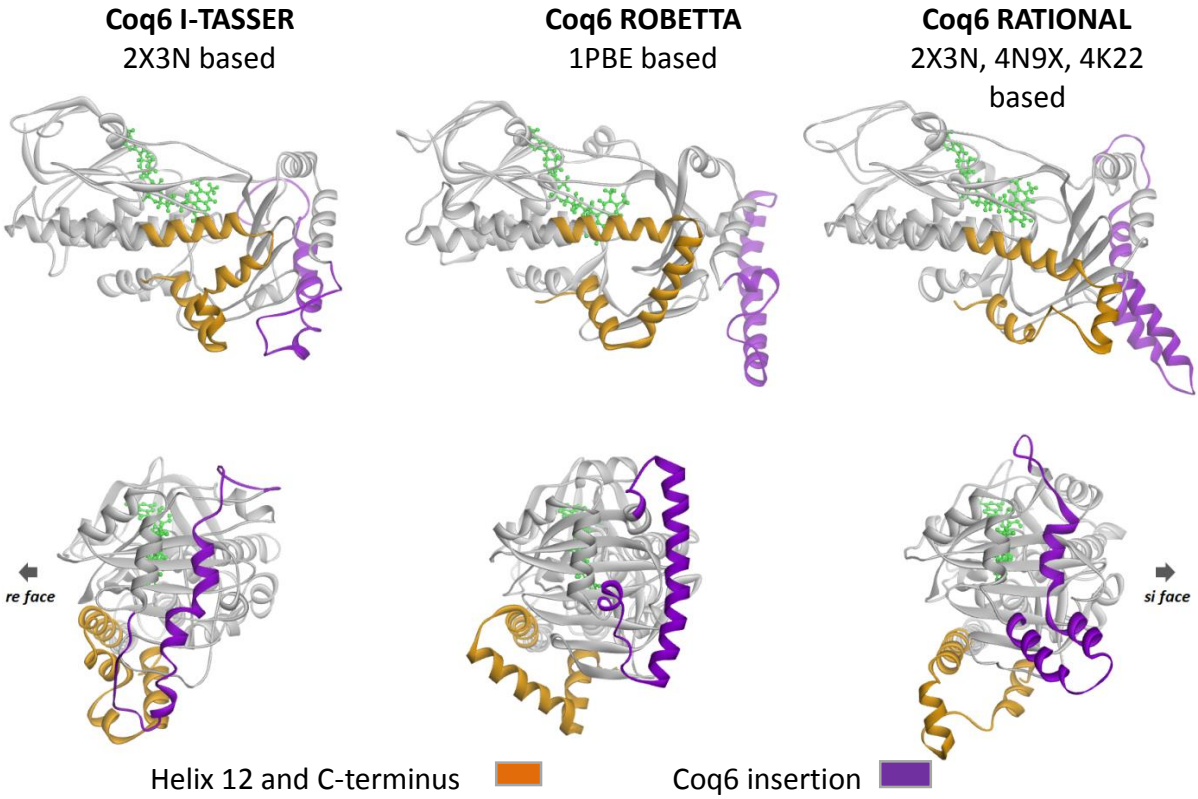

FIG 3.11 *Comparison of the three Coq6 homology models. The secondary structure is shown in cartoon and FAD is shown in green stick. The C-terminus and Coq6 family-insert are colored in orange and purple, respectively. From left to right: the I-TASSER model, the ROBETTA model, and the RATIONAL model. The top row shows models from the re face, while the bottom row is rotated 90 degrees about the vertical to show the exterior of the beta sheet domain, site of the insert.*

This set of full length wild-type Coq6 models, comprised of the Generation 3 RATIONAL model, the I-TASSER model, and the ROBETTA model will be subjected to molecular dynamics in order to test the stability, and therefore physical plausibility, of each model.

This set of models represents different possible conformations for the C-terminus as well as the Coq6-family insert. The ROBETTA model reproduces the large equilateral triangle of the C-terminus observed in its 1PBE template. The RATIONAL model reproduces the C-terminal geometry of 4N9X. The I-TASSER model brings a new C-terminal conformational possibility distinct from those generated by the other methods. The three models also propose conformations for the Coq6 insert. All are largely helical. ROBETTA constructs a single long helix connected to the protein by loosely structured coils. I-TASSER constructs a less orderly helix for the C-terminal half of the insert, while the N-terminal half of the insert runs in a loose antiparallel coil. This C-terminal half of the insert is structured as a helix in the multi-template model through the application of secondary structure restraints derived from the Jpred secondary structure prediction method.[29]

4. Molecular dynamics simulation of Generation 3 constructs

The I-TASSER, ROBETTA, and RATIONAL models of the full-length wild-type Coq6 enzyme were subjected to molecular dynamics in order to assess their structural stability. The protein models were solvated using TIP3P[30] water under the AMBER99SB-ILDN[31] force-field using PME[32] electrostatics in GROMACS 4.6.5.[33] Protonation states of titratable residues were assigned with PropKa[34] and the GROMACS pdb2gmx function. The salt concentration was set to 0.157M NaCl as documented for the mitochondrial matrix[35] The simulation cell was a rhombic dodecahedron allowing 1.4 nanometers between the protein and the cell edge. Models were subjected to 300 000 steps of steepest descent minimization. Equilibration was conducted in two phases (NVT and NPT) of 250 ps each at a time step of 1fs with position restraints on protein heavy atoms using the velocity-rescaled Berendsen thermostat[36] at 300K. NPT equilibration used the Parinello-Rahman barostat.[37] Bond lengths were not constrained during equilibration. FAD was imported from PHBH structure 1PBE and minimized after binding site rotamer adjustment. FAD force field parameters were provided by Sengupta *et al*.[38] The General Amber Force Field (GAFF)[39] procedure with partial charges optimized at the 6-31G* level was employed for generating this data. A representative FAD conformation from MD of the RATIONAL model is shown below in **Figure 3.12**. Production simulations for structural stability screening were run for 20ns using a time-step of 2fs, leap-frog Verlet, and LINCS constraints for heavy atom-hydrogen bonds. The temperatures of the protein and solvent were coupled separately to the Berendsen thermostat with a relaxation time of 0.1 ps at 300K. In total, three replicas for each model's simulation were run.

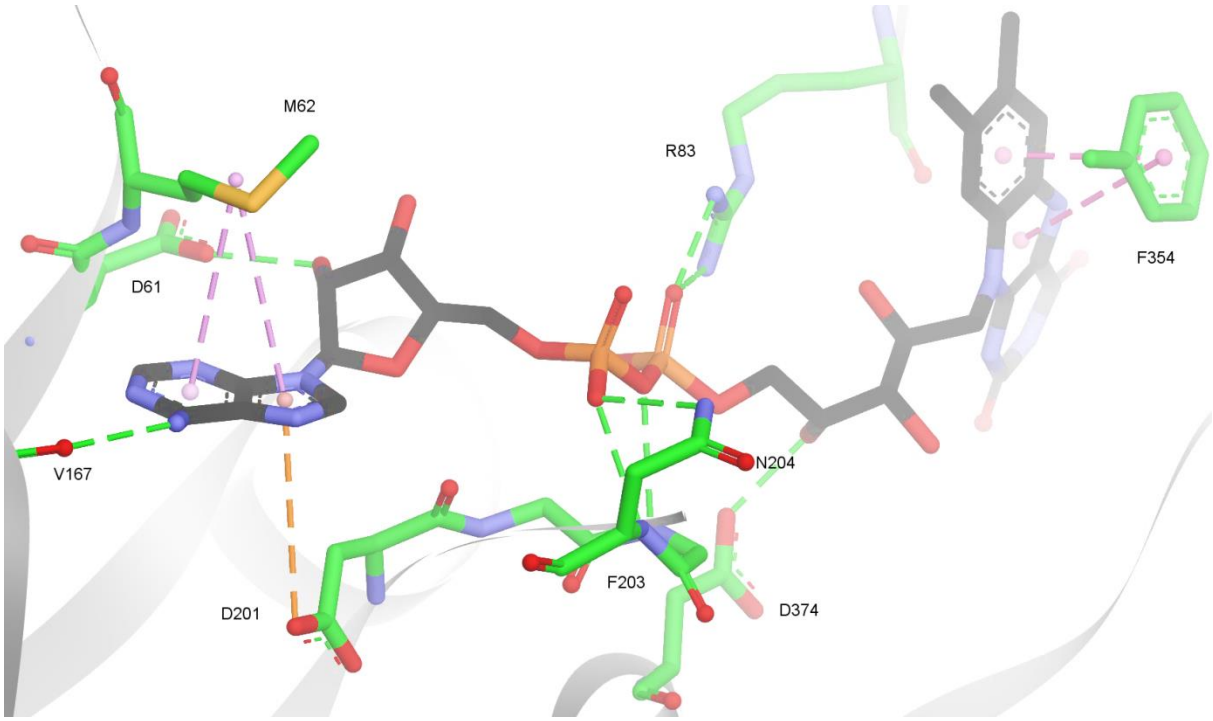

FIG 3.12 *Model of FAD in the Coq6 FAD binding site, as shown in the RATIONAL model. This pose is from a substrate binding conformation selected from MD. The adenine ring's N3 nitrogen is hydrogen bonded to V167. The adenine aromatic system can also form pi-sulfur interactions with M62 and pi-anion interactions with D201. The ribose hydroxyl is H bonded to D61. The pyrophosphate is H-bonded to the side chains of R83 and N204, as well as the backbone nitrogen of F203 (side chain omitted for clarity). These residues are highly conserved in the Coq6 family (as calculated with ConSurf), with the exception of M62 and F203.*

The primary purpose of the molecular dynamics screening of this model set is to evaluate the structural stability of each model, as **stability is a necessary but not sufficient pre-requisite for functionality** (which is explored later through other calculations). This is an important distinction **because models based on catalytically inactive templates may still be dynamically stable**. This distinction must be kept in mind primarily for the N-terminal region of the models because it comprises the majority of the protein and has many well-formed elements of secondary and super-secondary structure likely to remain stable despite lack of catalytic activity. However, the C-terminal region of the models is much smaller and has many fewer long- range contacts with the rest of the protein, making it more dependent on being properly structured intrinsically in order to be stable. Therefore, when reviewing the trajectories of the I-TASSER, ROBETTA, and RATIONAL models we will focus on the behavior of the C-terminus. The structural stability of the C-terminus can be described by how well it retains its secondary structure. Secondary structure can be calculated on the basis of the protein's internal coordinates (the backbone angles *phi* and *psi)* and it is a more accurate and robust description of local structural stability than atomic RMSD. This is because of the inherently non-directional averaging of atomic coordinates relative to the reference coordinates in the final computed RMSD value.

Two extreme cases of RMSD being an inaccurate descriptor of local structural stability exist. One is the rigid body displacement of a structure from its initial position. The conformation of the structure itself may be perfectly rigid, but an RMSD curve calculated based on its initial position will show a continuous increase, implying a structural deviation where there has only been a translational one. The other case is a conformational denaturation distributed evenly over the entire initial structure held at a fixed center of mass. In this case the RMSD curve calculated over time will be relatively flat, implying structural stability, while the conformation has changed significantly.

Therefore, to give a synoptic review of the MD simulations for this set we will show the first and last frames from the trajectory as well as the secondary structure description of the C-terminal region computed over the duration of the trajectories.

The secondary structure was visualized with the Timeline plugin for VMD.[40] The amino acid sequence of Coq6 is listed on the vertical axis, from the N-terminus at the top to the C-terminus at the bottom. The main structural domains of Coq6 are indicated by brackets. The horizontal axis shows time, from 0 to 20 nanoseconds. The secondary structure of each residue is assessed through the measure of protein backbone *phi* and *psi* angles, and is illustrated over time with a color scheme. Turns are green, strands are yellow, alpha helices are purple, 3-10 helices are blue, and coils are white. Stretches of residues maintaining a given secondary structure are visible as continuous horizontal bands of a given color.

4.1 I-TASSER Coq6 model 2X3N based

In this model the N-terminus has been modeled after 2X3N. The C-terminus has been modeled using I-TASSER's own protocol as a small 3 helix bundle. However, this structure is regionally unstable, with the helices dropping away from the bottom of the active site, as visible from comparing the first and last frames of dynamics (shown in **Figure 3.13**). In addition to moving far from their initial positions, they also display instabilities in their secondary structure, as can be seen in **Figure 3.14**. The C-terminus, which is initially has a high helical content (visible as horizontal purple bands at the bottom of the graphic) becomes interrupted by regions of green and blue, indicating that the secondary structure of this region is unstable. However, the rest of the model appears very stable so we will retain it for further comparative analysis.

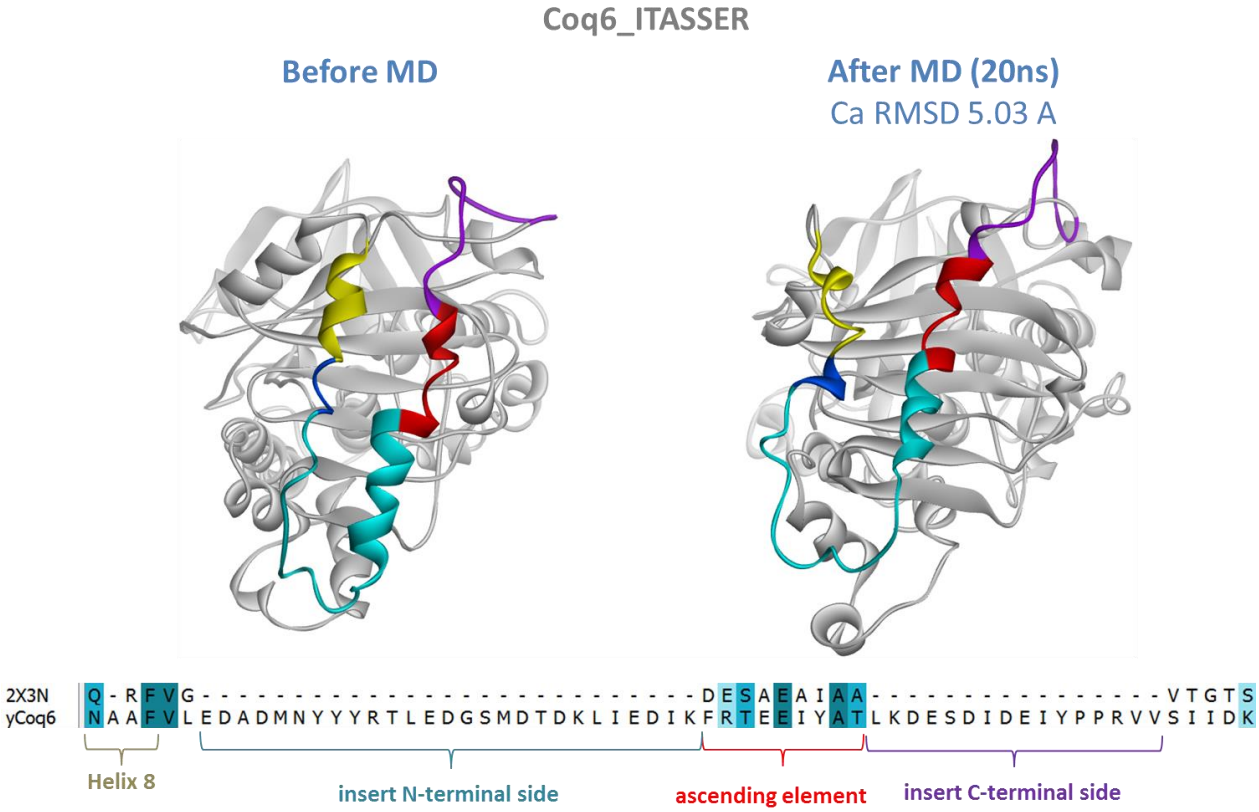

FIG 3.13 *I-TASSER Coq6 model, based on 2X3N: focus on the insert. A) before and B) after dynamics. The subsequences of the insert are color coded according to the legend.*

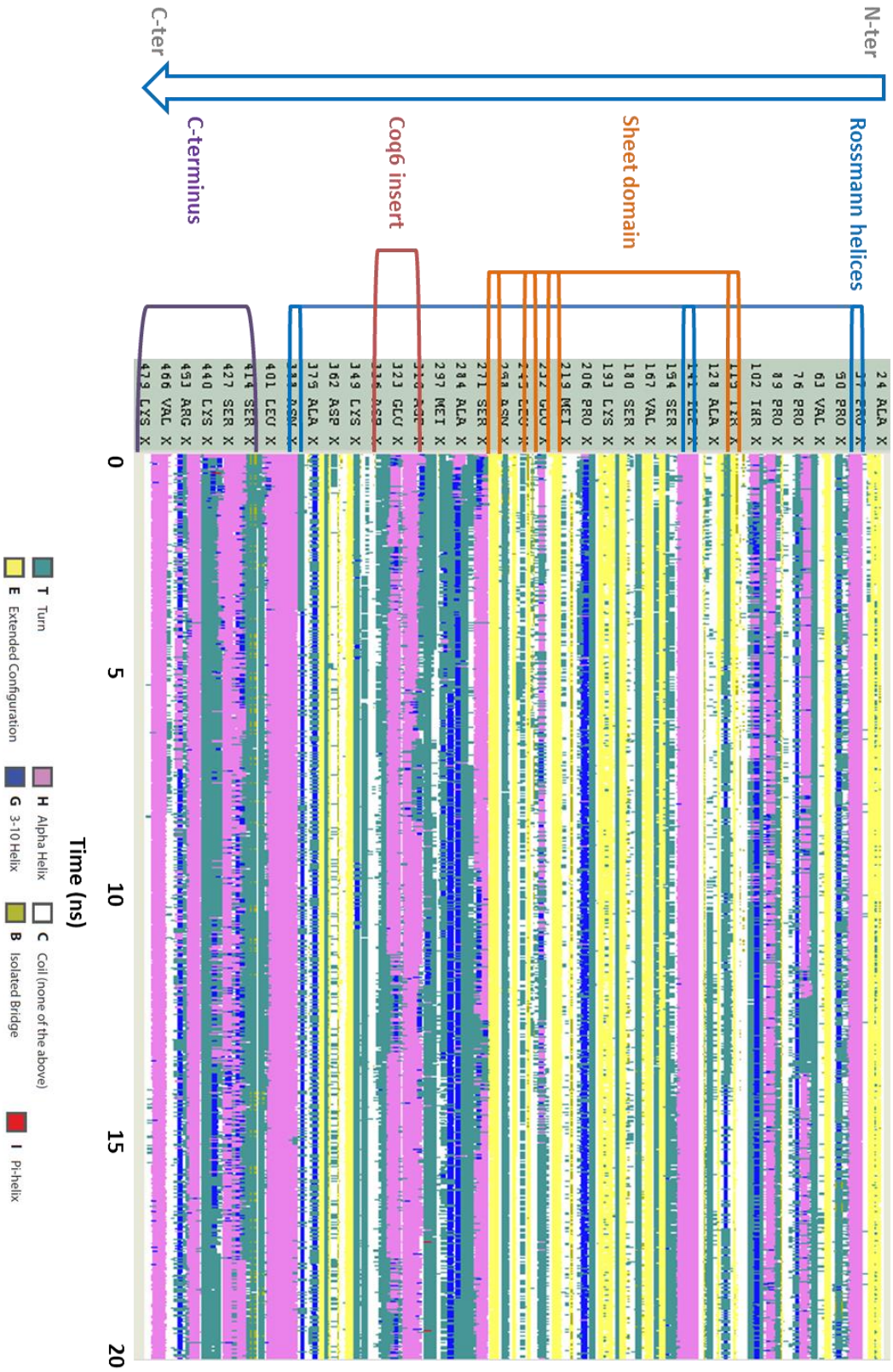

FIG 3.14 *I-TASSER Coq6 model, based on 2X3N: molecular dynamics stability screening review. Secondary structure persistence plot as calculated by the Timeline plugin for VMD over 20ns of simulation.*

The insert shows moderate stability. The regions initially reconstructed as helical remain largely helical, although they lack a well-defined secondary structure for the N-terminal side of the insert. Indeed, the secondary structure element immediately preceding the insert, Helix 8 (colored in gold **Figure 3.13**) is distorted compared to its initial position. This may reflect a region of the protein that is has less native or intrinsic structure, which is plausible if we consider that the insert's likely function is protein-protein association. It may not assume a well-defined structure in the absence of its protein partner. It is also possible that it is a naturally more mobile region, as enzymes of this class sometimes have regions too mobile to be crystallographically resolved, typically the *si*-loop.

4.2 ROBETTA Coq6 model 1PBE based

In this model the N-terminus has been modeled after 1PBE. The C-terminus has been modeled after the 1PBE structure as a large equilateral triangle. A 1PBE C-terminus mated to a 1PBE N-terminus gives a structure that is stable of 20ns, with the terminal helices showing minor movements. The C-terminus actually becomes *more* stable over the course of the simulation, as **Figure 3.16** shows the presence of 3-10 helices which transition to alpha helices over the second half of the simulation. This is also visible in the 3D structure before and after dynamics (as shown in **Figure 3.15**).

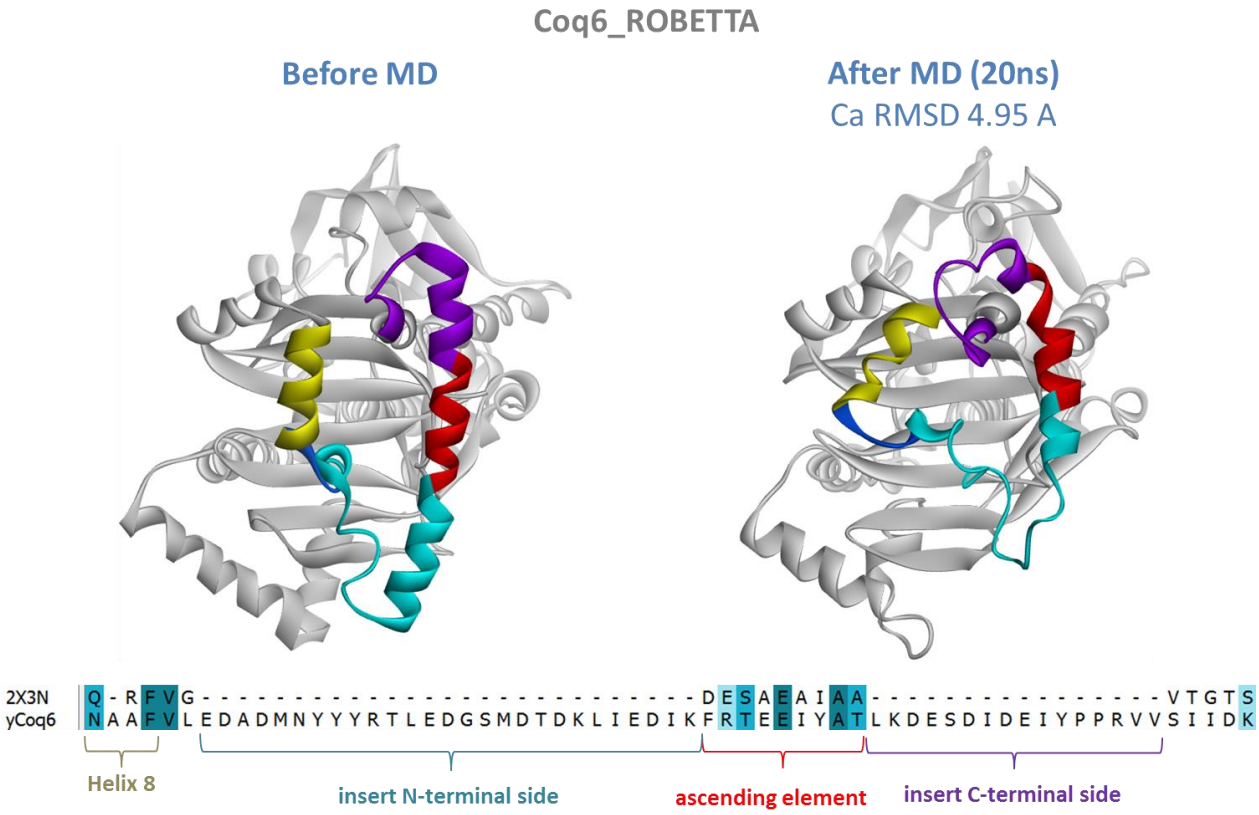

FIG 3.15 *ROBETTA Coq6 model, based on 1PBE: focus on the insert. A) before and B) after dynamics. The subsequences of the insert are color coded according to the legend.*

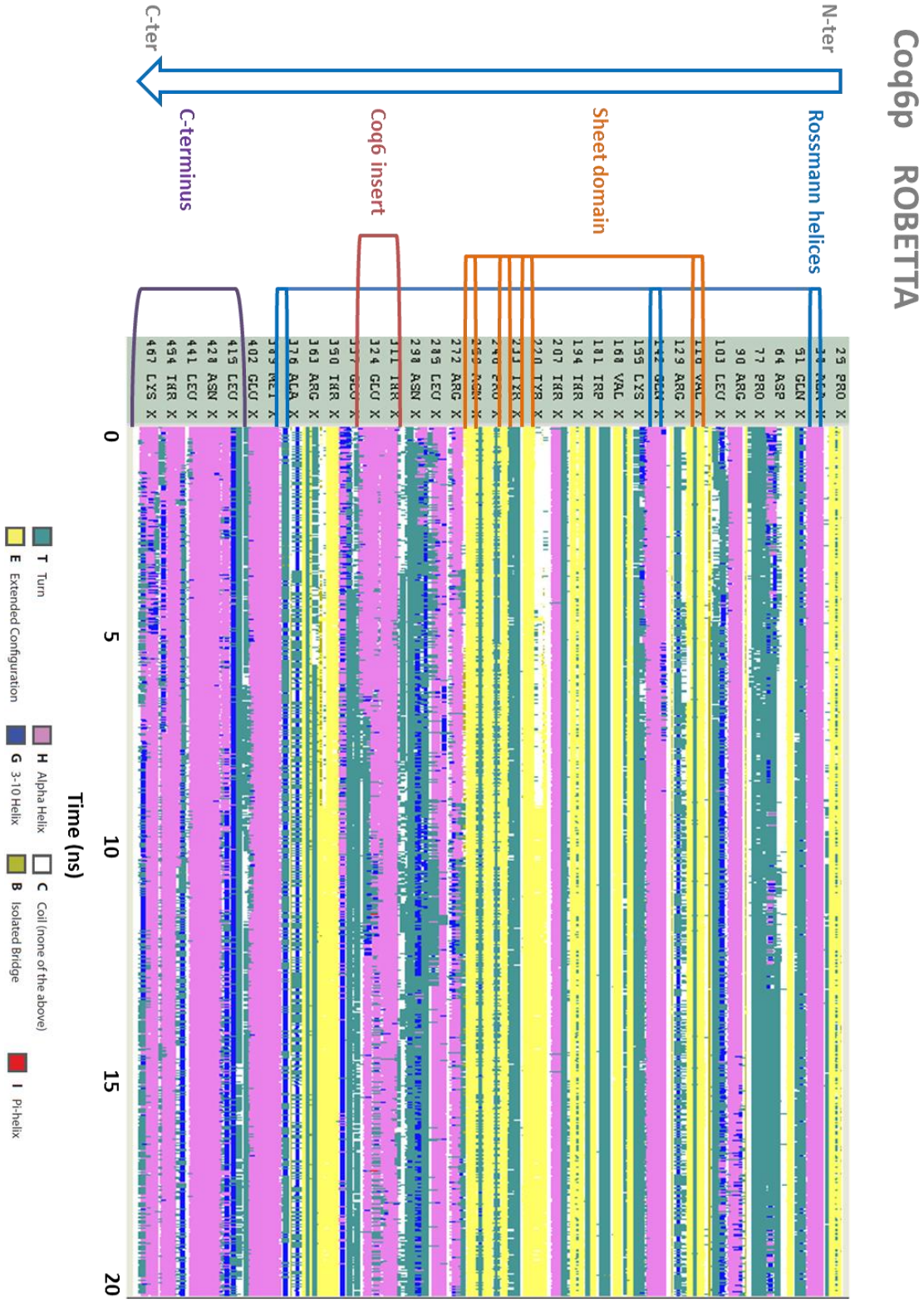

FIG 3.16 *ROBETTA Coq6 model, based on 1PBE: molecular dynamics stability screening review. Secondary structure persistence plot as calculated by the Timeline plugin for VMD over 20ns of simulation.*

In contrast, the ROBETTA structured insert shows less stability than in the I-TASSER model, deviating more from its initial conformation. The core of the helix (composed of the ascending element) is rather stable, maintaining its helicity until the end of the simulation. However, the helical structuring of the N- and C-terminal sides of the insert appear much less stable, losing their helicity and even distorting Helix 8 away from its original position. This indicates that the a mis-structured insert can destabilize proximal regions of the protein. While the general structure of the insert may involve a helix-turn-helix motif, its detailed secondary and tertiary structure is important to correctly define and position in model building and simulation.

4.3 RATIONAL Coq6 model 2X3N, 4N9X, and 4K22 based

In this model the N-terminus has been modeled after 2X3N. The C-terminus has been modeled after the 4N9X structure as an extended triangle. The C-terminus is notable for becoming *more* stable after dynamics, with secondary structures forming stable alpha helices over the second half of the simulation, as shown in the bottom portion of **Figure 3.18**, visible as the formation of horizontal purple stripes.

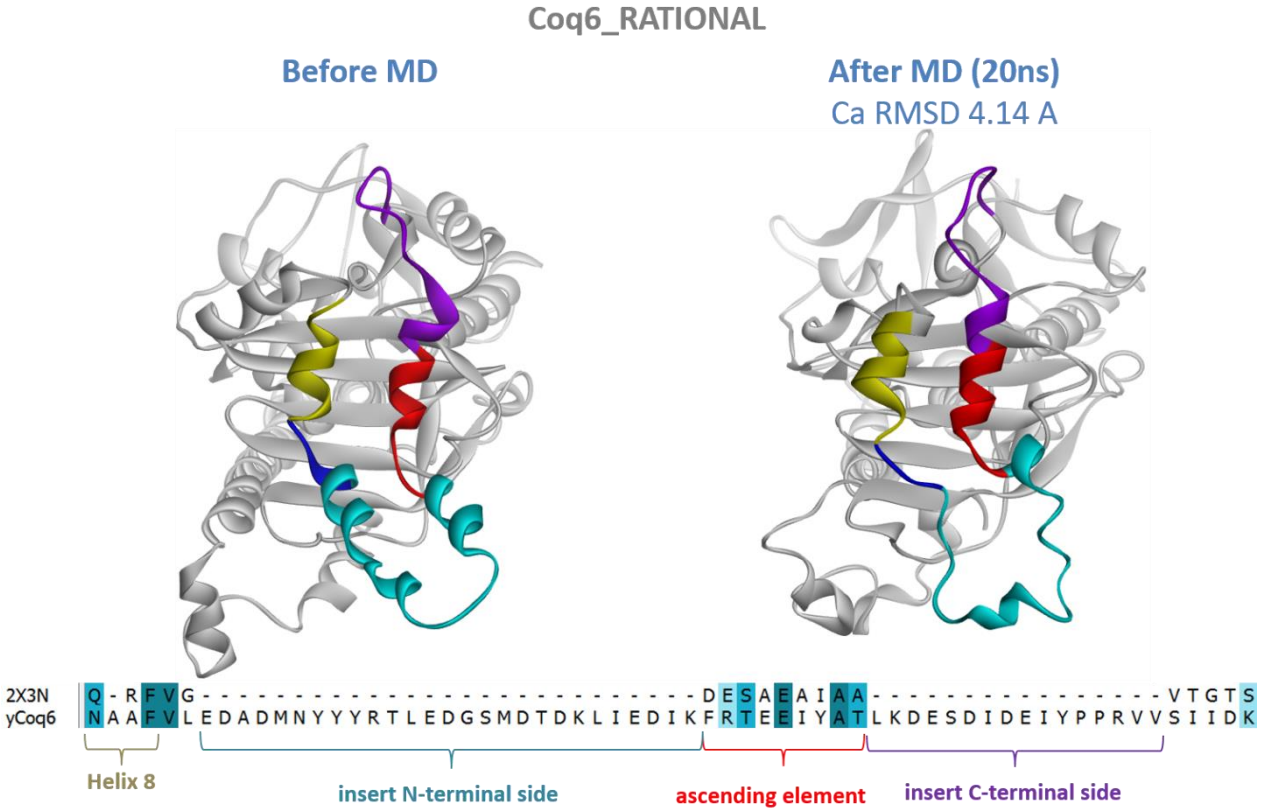

FIG 3.17 *RATIONAL Coq6 model, based on 2X3N and 4N9X: focus on the insert. A) before and B) after dynamics. The subsequences of the insert are color coded according to the legend.*

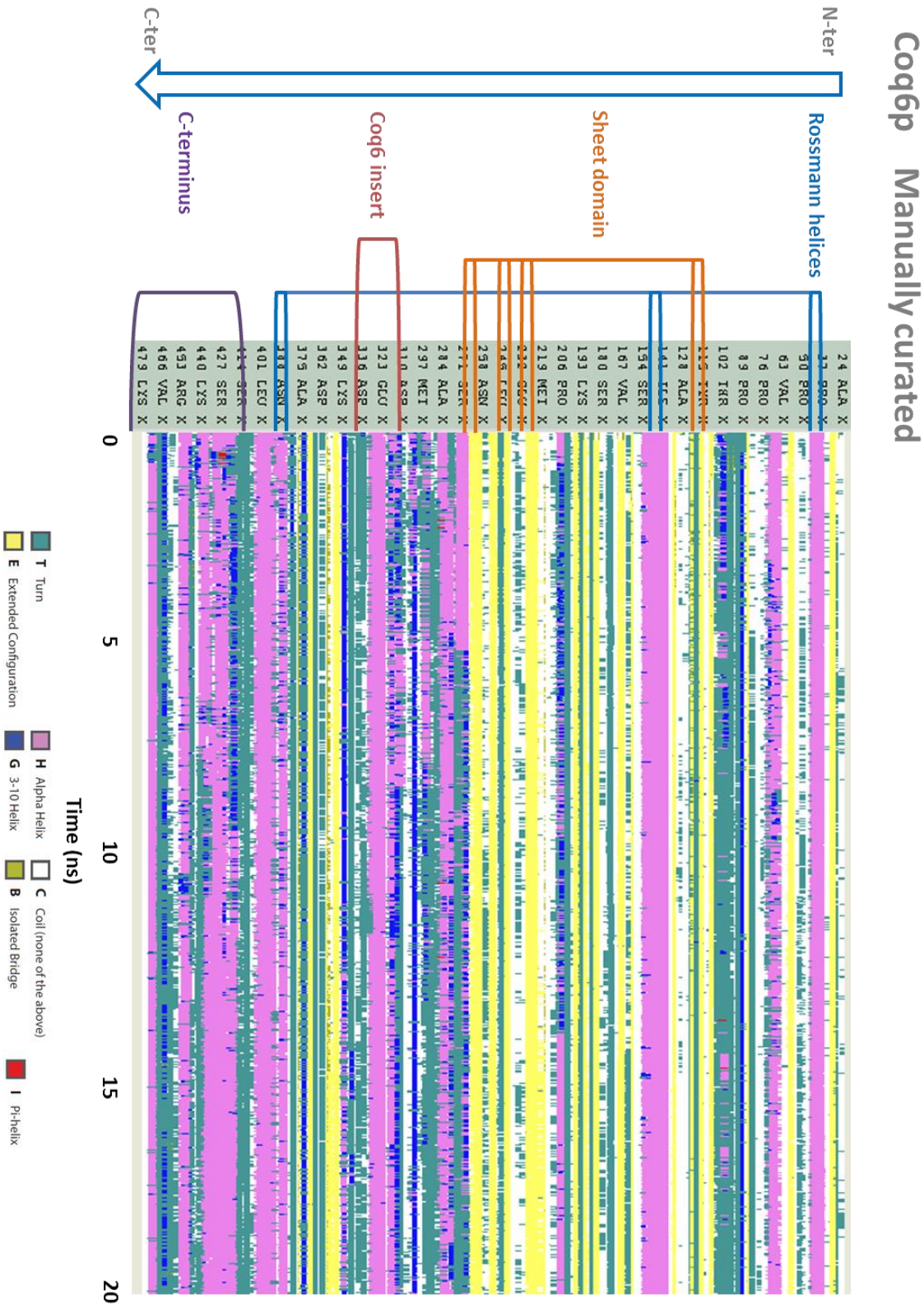

FIG 3.18 *RATIONAL Coq6 model, based on 2X3N and 4N9X: molecular dynamics stability screening review. Secondary structure persistence plot as calculated by the Timeline plugin for VMD over 20ns of simulation.*

The insert of the RATIONAL model shows some general similarities with the structure proposed by I-TASSER. The C-terminal side of the insert, including the ascending elements, retains most of its helicity. However, the N-terminal side of the insert, despite its initial helical structure, loses helicity as well, becoming a loose coil similar to the I-TASSER construction. However, despite the loosening of the insert's initial structure, it does not distort Helix 8 (colored in gold in **Figure 3.17**) or the ascending element (colored in red and purple in **Figure 3.17**).

4.4 Comparative regional RMSD summary plots

The goal of the Generation 3 models is to have molecular models of the full-length wild-type Coq6 enzyme complexed with FAD for the eventual purpose of substrate docking calculations. Therefore, we will analyze these models' stability more carefully. Since each of these models contain significant regions that were modeled from different coordinate sources, we decided to compute the RMSD for each major structural region as another way to compare their stability. We partitioned the models into 5 regions: the Rossmann-fold helices, the beta-sheet domain, and FAD binding site, the insert, and the C-terminus. We also monitor the global RMSD of the entire protein. The time evolution of the RMSDs is presented in the **Figure 3.19** below.

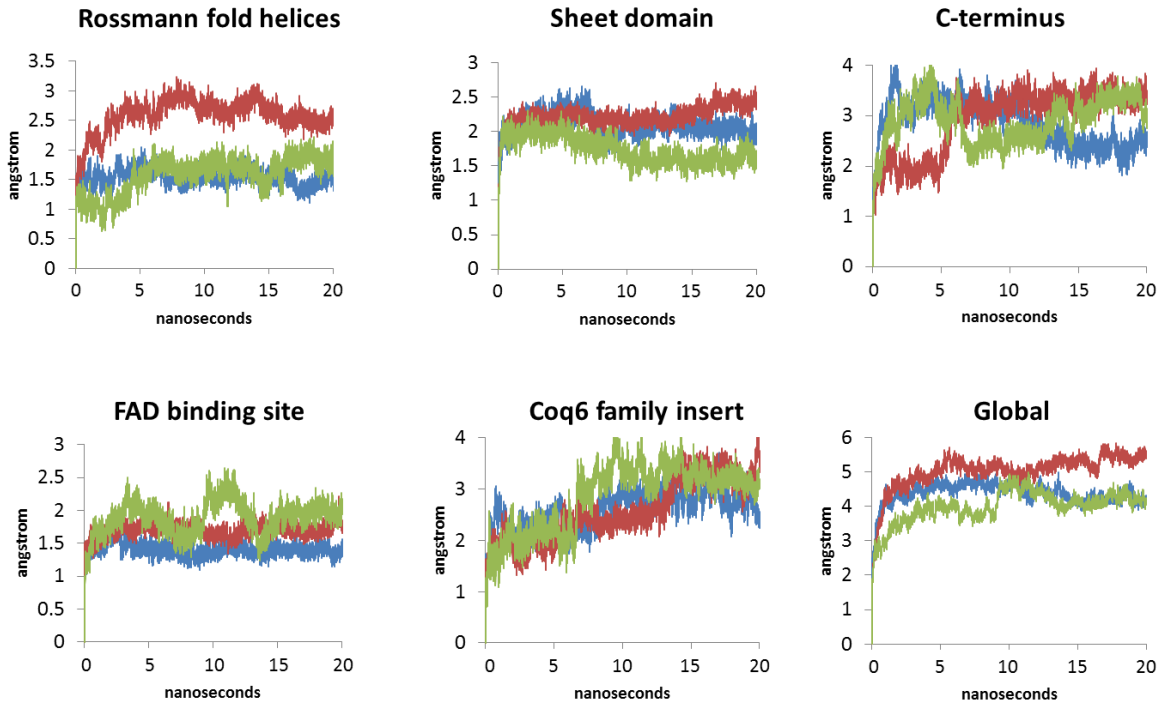

FIG 3.19 *Time evolution of RMSDs for selected regions of Coq6 models. The I-TASSER model RMSDs are shown in red; ROBETTA model RMSDs are shown in green; RATIONAL model RMSDs are shown in blue.*

The Rossmann helices form the core of the enzyme and appear relatively stable, with all three models reaching RMSD plateaus in this region for the latter half of the simulation. The sheet domain appears to be of roughly equal stability among all three models. These regions are common to all templates and

well defined crystallographically. I-TASSER's C-terminus shows a large shape change early in the simulation which stabilizes at a relatively high value by the end, similar to the ROBETTA C-terminus. The RATIONAL model C-terminus behaves differently, with RMSD decreasing over time after an initial jump, suggesting the attainment of a more stable conformation than the other models. Interestingly, the same is also true for the FAD binding site, with the RATIONAL model showing the least deviation from its initial conformation over the simulation, noticeably less than for the other models. In all models the insert structure is of questionable stability, as its RMSD seems to increase up until the end of the simulation. However, this is not critical for our results, since the insert is likely to participate in protein-protein interactions and may not assume a stable structure without its partners.

We can note some general trends: The I-TASSER model generally shows the highest RMSD values for each structural region, with the exception of the FAD binding site and the Coq6 insert, which show maximum mobility in the ROBETTA model. Higher RMSDs for these regions indicate they are less structurally stable than the RATIONAL model, which generally shows the lowest RMSD values during the latter half of the simulation. The combined effects of all these behaviors is even visible in the plot of global RMSD. The I-TASSER model shows an ever increasing RMSD at the end of the simulation, whereas the ROBETTA model shows some distinct metastable states before beginning to settle into a plateau at the end. The RATIONAL model's global RMSD is interesting because of how early it begins to decrease an converge towards a similar final value as the ROBETTA model.

5. Conclusion

The results of MD stability testing are consistent with what is physically known about each of the templates used to construct them. The first generation of homology models without the insert, based on Coq6 alignments against 4K22, had to be discarded because of their unrealistic geometry in the vicinity of the insert and the mal-formed FAD binding GDAxH loop. The second generation of rationally designed models, based largely on 2X3N and without the insert allowed us to test the stability of different combinations of N- and C-terminal templates through MD, revealing the combination of 2X3N and 4N9X templates to be optimal. Templates using 4K22 for the N-terminus or 1PBE for the C-terminus show instability in the C-terminal region. This is not surprising, since the coordinates of 4K22 describe a physical system in the absence of a C-terminus (as well as without FAD).

The third generation of rationally designed models based on 2X3N and 4N9X *with* the insert finally allowed us to test the structure of the insert in our own RATIONAL model, and explore the conformations of additional Coq6 models independently generated by the automated modeling servers I-TASSER an ROBETTA. Molecular dynamics reveals that all three models are structurally stable enough for further calculations. Therefore, only these three models will be retained for the next steps: an analysis of the models to find and test possible substrate binding sites of the Coq6 enzyme.

References

[1] Sánchez, Roberto, and Andrej Šali. “Comparative Protein Structure Modeling as an Optimization Problem.” *Journal of Molecular Structure: THEOCHEM*, World Congress of Theoretically Oriented Chemists, 398–99 (June 30, 1997): 489–96. doi:10.1016/S0166-1280(96)04971-8.

[2] Lawrence, Lawrence A. Kelley, and Michael J. E. Sternberg. “Protein Structure Prediction on the Web: A Case Study Using the Phyre Server.” *Nature Protocols* 4, no. 3 (February 2009): 363–71. doi:10.1038/nprot.2009.2.

[3] Schreuder, Herman A., Peter A. J. Prick, Rik K. Wierenga, Gerrit Vriend, Keith S. Wilson, Wim G. J. Hol, and Jan Drenth. “Crystal Structure of the P-Hydroxybenzoate Hydroxylase-Substrate Complex Refined at 1.9 Å Resolution: Analysis of the Enzyme-Substrate and Enzyme-Product Complexes.” *Journal of Molecular Biology* 208, no. 4 (August 20, 1989): 679–96. doi:10.1016/0022-2836(89)90158-7.

[4] Illergård, Kristoffer, David H. Ardell, and Arne Elofsson. “Structure Is Three to Ten Times More Conserved than Sequence--a Study of Structural Response in Protein Cores.” *Proteins* 77, no. 3 (November 15, 2009): 499–508. doi:10.1002/prot.22458.

[5] Chehade, Mahmoud Hajj, Laurent Loiseau, Murielle Lombard, Ludovic Pecqueur, Alexandre Ismail, Myriam Smadja, Béatrice Golinelli-Pimpaneau, et al. “ubiI, a New Gene in Escherichia Coli Coenzyme Q Biosynthesis, Is Involved in Aerobic C5-Hydroxylation.” *Journal of Biological Chemistry* 288, no. 27 (July 5, 2013): 20085–92. doi:10.1074/jbc.M113.480368.

[6] Kuzin, A., Y. Chen, S. Lew, J. Seetharaman, L. Mao, R. Xiao, L. A. Owens, et al. “Crystal Structure of the OCTAPRENYL-METHYL-METHOXY-BENZQ MOLECULE from Erwina Carotovora Subsp. Atroseptica Strain SCRI 1043 / ATCC BAA-672, Northeast Structural Genomics Consortium (NESG) Target EwR161.” *To Be Published*, October 21, 2013, null – null.

[7] Holm, Liisa, and Päivi Rosenström. “Dali Server: Conservation Mapping in 3D.” *Nucleic Acids Research* 38, no. suppl 2 (July 1, 2010): W545–49. doi:10.1093/nar/gkq366.

[8] Oke, M., L. G. Carter, K. A. Johnson, H. Liu, S. A. Mcmahon, X. Yan, M. Kerou, et al. “2X3N - The Scottish Structural Proteomics Facility: Targets, Methods and Outputs.” *J.Struct.Funct.Genomics* 11 (January 25, 2010): 167–80. doi:10.1007/s10969-010-9090-y.

[9] Eppink, M. H., H. A. Schreuder, and W. J. van Berkel. “Lys42 and Ser42 Variants of P-Hydroxybenzoate Hydroxylase from Pseudomonas Fluorescens Reveal That Arg42 Is Essential for NADPH Binding.” *Eur.J.Biochem.* 253 (May 26, 1998): 194–201. doi:10.1046/j.1432-1327.1998.2530194.x.

[10] Buedenbender, S., S. Rachid, R. Muller, and G. E. Schulz. “Structure and Action of the Myxobacterial Chondrochloren Halogenase CndH: A New Variant of FAD-Dependent Halogenases.” *J.Mol.Biol.* 385 (August 4, 2008): 520–30. doi:10.1016/j.jmb.2008.10.057.

[11] Gatti, D. L., B. A. Palfey, M. S. Lah, B. Entsch, V. Massey, D. P. Ballou, and M. L. Ludwig. “The Mobile Flavin of 4-OH Benzoate Hydroxylase.” *Science* 266 (September 6, 1994): 110–14.

[12] Vorobiev, S., M. Su, J. Seetharaman, S. Sahdev, R. Xiao, E. L. Foote, C. Ciccosanti, et al. “Crystal Structure of Flavoprotein/dehydrogenase from Cytophaga Hutchinsonii.” *To Be Published*, June 16, 2010, null – null.

[13] Cole, Lindsay J., Barrie Entsch, Mariliz Ortiz-Maldonado, and David P. Ballou. “Properties of P-Hydroxybenzoate Hydroxylase When Stabilized in Its Open Conformation.” *Biochemistry* 44, no. 45 (November 15, 2005): 14807–17. doi:10.1021/bi0512142.

[14] Fox, Naomi K., Steven E. Brenner, and John-Marc Chandonia. “SCOPe: Structural Classification of Proteins—extended, Integrating SCOP and ASTRAL Data and Classification of New Structures.” *Nucleic Acids Research* 42, no. D1 (January 1, 2014): D304–9. doi:10.1093/nar/gkt1240.

[15] van Berkel, W. J. H., N. M. Kamerbeek, and M. W. Fraaije. “Flavoprotein Monooxygenases, a Diverse Class of Oxidative Biocatalysts.” *Journal of Biotechnology* 124, no. 4 (August 5, 2006): 670–89. doi:10.1016/j.jbiotec.2006.03.044.

[16] Crozier-Reabe, Karen, and Graham R. Moran. “Form Follows Function: Structural and Catalytic Variation in the Class A Flavoprotein Monooxygenases.” *International Journal of Molecular Sciences* 13, no. 12 (November 23, 2012): 15601–39. doi:10.3390/ijms131215601.

[17] Huse, Holly, and Marvin Whiteley. “4-Quinolones: Smart Phones of the Microbial World.” *Chemical Reviews* 111, no. 1 (January 12, 2011): 152–59. doi:10.1021/cr100063u.

[18] Schreuder, H. A., J. M. van der Laan, W. G. Hol, and J. Drenth. “Crystal Structure of P-Hydroxybenzoate Hydroxylase Complexed with Its Reaction Product 3,4-Dihydroxybenzoate.” *Journal of Molecular Biology* 199, no. 4 (February 20, 1988): 637–48.

[19] Schreuder, Herman A., Peter A. J. Prick, Rik K. Wierenga, Gerrit Vriend, Keith S. Wilson, Wim G. J. Hol, and Jan Drenth. “Crystal Structure of the P-Hydroxybenzoate Hydroxylase-Substrate Complex Refined at 1.9 Å Resolution: Analysis of the Enzyme-Substrate and Enzyme-Product Complexes.” *Journal of Molecular Biology* 208, no. 4 (August 20, 1989): 679–96. doi:10.1016/0022-2836(89)90158-7.

[20] S. Montersino, D. Tischler. “Catalytic and Structural Features of Flavoprotein Hydroxylases and Epoxidases.” *Advanced Synthesis & Catalysis* 353 (2011): 2301–19.

[21] Gin, Peter, Adam Y. Hsu, Steven C. Rothman, Tanya Jonassen, Peter T. Lee, Alexander Tzagoloff, and Catherine F. Clarke. “The Saccharomyces Cerevisiae COQ6 Gene Encodes a Mitochondrial Flavin-Dependent Monooxygenase Required for Coenzyme Q Biosynthesis.” *Journal of Biological Chemistry* 278, no. 28 (July 11, 2003): 25308–16. doi:10.1074/jbc.M303234200.

[22] Celniker, Gershon, Guy Nimrod, Haim Ashkenazy, Fabian Glaser, Eric Martz, Itay Mayrose, Tal Pupko, and Nir Ben-Tal. "ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function." *Israel Journal of Chemistry* 53, no. 3–4 (April 1, 2013): 199–206. doi:10.1002/ijch.201200096.

[23] Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. "Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7, no. 1 (January 1, 2011): 539. doi:10.1038/msb.2011.75.

[24] Drozdetskiy, Alexey, Christian Cole, James Procter, and Geoffrey J. Barton. "JPred4: A Protein Secondary Structure Prediction Server." *Nucleic Acids Research*, April 16, 2015, gkv332. doi:10.1093/nar/gkv332.

[25] He, Cuiwen H., Letian X. Xie, Christopher M. Allan, UyenPhuong C. Tran, and Catherine F. Clarke. "Coenzyme Q Supplementation or over-Expression of the Yeast Coq8 Putative Kinase Stabilizes Multi-Subunit Coq Polypeptide Complexes in Yeast Coq Null Mutants." *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1841, no. 4 (April 2014): 630–44. doi:10.1016/j.bbalip.2013.12.017.

[26] Zhang, Yang. "I-TASSER Server for Protein 3D Structure Prediction." *BMC Bioinformatics* 9, no. 1 (January 23, 2008): 40. doi:10.1186/1471-2105-9-40.

[27] Song, Yifan, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, TJ Brunette, James Thompson, and David Baker. "High-Resolution Comparative Modeling with RosettaCM." *Structure* 21, no. 10 (October 8, 2013): 1735–42. doi:10.1016/j.str.2013.08.005.

[28] Kabsch, Wolfgang, and Christian Sander. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features." *Biopolymers* 22, no. 12 (December 1, 1983): 2577–2637. doi:10.1002/bip.360221211.

[29] Drozdetskiy, Alexey, Christian Cole, James Procter, and Geoffrey J. Barton. "JPred4: A Protein Secondary Structure Prediction Server." *Nucleic Acids Research*, April 16, 2015, gkv332. doi:10.1093/nar/gkv332.

[30] Jorgensen, William L., Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. "Comparison of Simple Potential Functions for Simulating Liquid Water." *The Journal of Chemical Physics* 79, no. 2 (July 15, 1983): 926–35. doi:10.1063/1.445869.

[31] Lindorff-Larsen, Kresten, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. "Improved Side-Chain Torsion Potentials for the Amber ff99SB Protein Force Field." *Proteins* 78, no. 8 (June 2010): 1950–58. doi:10.1002/prot.22711.

[32] Essmann, Ulrich, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. "A Smooth Particle Mesh Ewald Method." *The Journal of Chemical Physics* 103, no. 19 (November 15, 1995): 8577–93. doi:10.1063/1.470117.

[33] Van Der Spoel, David, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. "GROMACS: Fast, Flexible, and Free." *Journal of Computational Chemistry* 26, no. 16 (December 1, 2005): 1701–18. doi:10.1002/jcc.20291.

[34] Li, Hui, Andrew D. Robertson, and Jan H. Jensen. "Very Fast Empirical Prediction and Rationalization of Protein pKa Values." *Proteins: Structure, Function, and Bioinformatics* 61, no. 4 (December 1, 2005): 704–21. doi:10.1002/prot.20660.

[35] Ballantyne, Js, and Cd Moyes. "The Effects of Salinity Acclimation on the Osmotic Properties of Mitochondria from the Gill of Crassostrea-Virginica." *Journal of Experimental Biology* 133 (November 1987): 449–56.

[36] Bussi, Giovanni, Davide Donadio, and Michele Parrinello. "Canonical Sampling through Velocity Rescaling." *The Journal of Chemical Physics* 126, no. 1 (January 7, 2007): 014101. doi:10.1063/1.2408420.

[37] Parrinello, M., and A. Rahman. "Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method." *Journal of Applied Physics* 52, no. 12 (December 1, 1981): 7182–90. doi:10.1063/1.328693.

[38] Sengupta, Abhigyan, Wilbee D. Sasikala, Arnab Mukherjee, and Partha Hazra. "Comparative Study of Flavins Binding with Human Serum Albumin: A Fluorometric, Thermodynamic, and Molecular Dynamics Approach." *ChemPhysChem* 13, no. 8 (June 4, 2012): 2142–53. doi:10.1002/cphc.201200044.

[39] Wang, Junmei, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. "Development and Testing of a General Amber Force Field." *Journal of Computational Chemistry* 25, no. 9 (July 15, 2004): 1157–74. doi:10.1002/jcc.20035.

[40] Humphrey, William, Andrew Dalke, and Klaus Schulten. "VMD: Visual Molecular Dynamics." *Journal of Molecular Graphics* 14, no. 1 (February 1996): 33–38. doi:10.1016/0263-7855(96)00018-5.

Chapter 4 Selection of Coq6 models through molecular dynamics and substrate docking

1. Introduction

The goal of our homology modeling is to create a model of Coq6 of sufficient accuracy to produce testable structure-function hypotheses, particularly for substrate binding. These hypotheses can be tested through *in vivo* activity assays of site-directed mutants. Therefore, we will subject our models to series of analyses designed to characterize substrate binding regions, including the active site region. This distinction is important because the large size and hydrophobicity of the Coq6 substrate (a hexaprenylated Q biosynthesis intermediate) is likely to require a substrate binding region significantly larger than what is necessary for just the aromatic head, which is the site of hydroxylation.

Thus far we have produced three alternative homology models of the Coq6 enzyme (I-TASSER, ROBETTA, and RATIONAL) which have satisfied the criterion of structural stability as assessed by molecular dynamics simulations. We now want to use these three models to attempt to make predictions about enzyme-substrate interactions. This is the domain of molecular docking.

2. Selection of Coq6 models by substrate docking

2.1 Receptor-ligand binding: induced fit vs. conformational selection

Receptor-ligand binding is the meeting of two molecules. Each molecule is a flexible object in constant motion, with distinct regional qualities such as polarity and hydrophobicity arising from electronic structure. This means that even a single isolated molecule is better represented as a collection of conformations rather than as a single static object. Thus, when two molecules meet, the process is better thought of as a meeting between these two collections. Some conformations of one molecule may be complementary to some conformations of the other molecule, and the actual binding event depends on the ability of the molecules to find these complementary conformations. The search for these complementary conformations is the domain of molecular docking, and dealing with highly flexible interacting objects is its central challenge.

A fundamental point in using molecular docking to investigate receptor-ligand binding is the question of how much the conformational distribution of one molecule affects the conformational distribution of the other. Does each molecule retain its own conformational distribution arising from its intrinsic geometry preferences during the binding event? Or do their distributions change in response to each other? This is also known as the question of *conformational selection* versus *induced fit*,[1] which are perhaps two extremes of a continuum of interaction types possible in the more general conceptual model of interacting conformational distributions.

The conformational selection model of ligand-receptor binding posits that the ligand bound conformation of the receptor pre-exists in the conformational space accessible to the receptor in the absence of the ligand, and that the presence of the ligand biases the conformational distribution of the receptor towards this binding conformation. The induced-fit model of binding states that the ligand-bound conformation of the receptor is not accessible in the conformational space of the receptor in the absence of the ligand (likely due to the presence of an energy barrier in the potential energy function which is lowered in the presence of the ligand), and that physical interaction with the ligand is required to produce the receptor's bound conformation. The distinction can essentially be re-stated as the ability

of the receptor to populate its ligand-bound conformation in the absence of the ligand. However, the ability of a given receptor to populate a ligand-binding conformational state it is not necessarily a binary condition: it is likely to be proportional to the barrier height to that local minimum, and barrier height can vary continuously. Therefore, a relatively low barrier to the ligand binding conformation may be crossed, albeit at low frequency, even without the ligand. Indeed, most real-world ligand-receptor systems are probably between the two conceptual extremes.

The practical implication of this distinction for molecular modeling studies of Coq6 is that in the case of pure conformational selection (at one extreme of the spectrum), the receptor's bound conformation may be accessible through conformational sampling of the ligand-free receptor. In our case, Coq6 conformations are sampled with molecular dynamics without the substrate. In the case of pure induced-fit, the receptor's binding conformation will never be accessible through MD conformational sampling without the substrate, necessitating its inclusion in the simulation. While it is possible to develop both types of simulations (without or with the substrate), the two types of simulations cannot be developed in parallel because the latter requires more underlying hypotheses. **Molecular dynamics simulations including the substrate will require coordinates for the initial position of the substrate, which are not known *a priori* by any method, experimentally or computationally**. **Therefore, the practical order for studying Coq6 substrate-receptor binding is to first study Coq6 without the substrate in order to determine likely substrate binding regions. This is why we will proceed with the simpler option first: simulation without the substrate, which corresponds to the hypothesis of conformational selection.**

A preliminary conformational study on one of the templates, PHBH co-crystallized with its substrate (para-hydroxybenzoate, pHB) suggests that conformational selection is likely to be applicable to the Coq6 system. Our molecular dynamics of simulations of PHBH show that the enzyme was able to sample the substrate-bound conformation without the substrate being present in the simulation. This was established by defining a receptor-based scoring function to describe the geometry of the substrate-bound active-site of PHBH. The interatomic distances defined in the scoring function were used as criteria for selecting conformations from molecular dynamics most similar to the co-crystal reference conformation. We were able to re-dock the substrate into these selected conformations and were able to reproduce the crystal pose, as shown in **Figure 4.1** below.

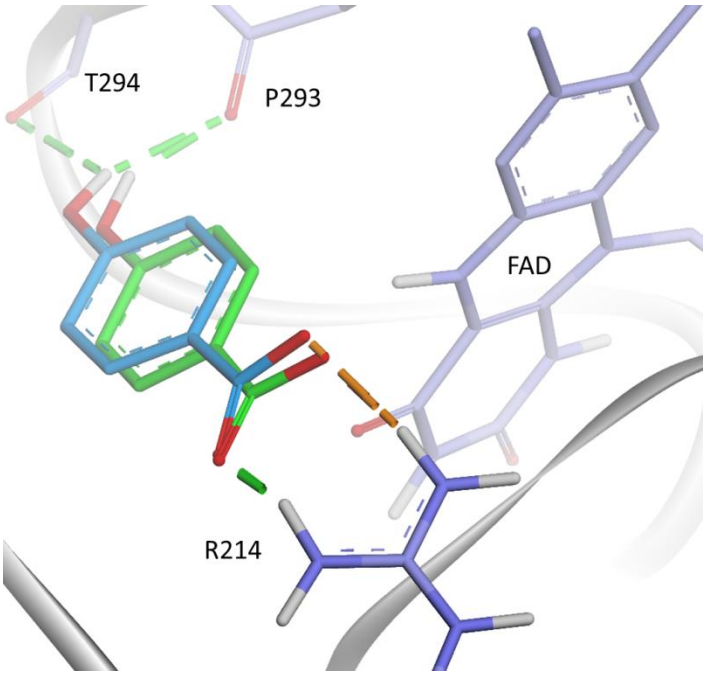

FIG 4.1: *Redocking of pHB into PHBH reproduces the crystal pose. The PHBH conformation was selected from molecular dynamics simulations using an active site geometry scoring function described in more detail in Section 5.2. Protein and FAD carbons are shown in lavender; the pHB from the 1PBE crystal structure is shown in green, while the docked pHB pose is in cyan.*

2.2 Receptor-ligand binding as approximated by ensemble docking

Having gathered support from PHBH that the receptor-ligand interaction in the Coq6 system is likely to be dominated by conformational selection, we still need a practical way to do molecular docking. Generally, docking algorithms represent a receptor as a static structure, only allowing the ligand to be flexible during the fitting process. This is because the computational expense of representing many degrees of freedom grows exponentially with the number of rotatable bonds. This computational limitation places a great importance on the specific conformation of the receptor used. If the specific coordinates of the receptor are not already in a position compatible with substrate binding, the docking process can easily fail to predict the correct receptor-ligand interactions – even if these interactions are known experimentally *a priori*. Thus, the use of a single rigid receptor structure is a serious limitation to the predictive ability of docking. Some algorithms (including AutoDock Vina) allow a small number of receptor sidechains to be flexible, but this typically does not allow any backbone movements, as might be implicated in the opening or closing of a substrate binding site. However, in real life both receptor and ligand are free to move, and indeed, are in continuous motion.

Another way of representing the conformational flexibility of the receptor is to generate many specific conformations of it, and attempt ligand docking into each static structure. The set of receptor conformations is often called an *ensemble*, and gives the technique its name: ensemble docking.[2] It seeks to address the representation of receptor flexibility through the testing of many individual receptor conformations. The next question is how to generate these alternative receptor conformations. Sometimes they are experimentally available, through the crystallographic resolution of multiple conformations, or through the conformational ensembles computed from NMR data. Since we have no

such structural data for Coq6, we must again turn to techniques of molecular modeling. Fortunately, molecular dynamics provides a way of computing ensembles of conformations. However, MD generates a lot of data: potentially thousands of conformations for even a short time period of 20 nanoseconds.

The question becomes how to select representative conformations from MD for substrate docking. We will develop this further in a following section. However, before proceeding to a more in depth study of ensemble docking, we will first evaluate our tool for molecular docking, an algorithm called AutoDock Vina.

2.3 Preliminary study on substrate models

Since molecular docking is a technique computationally limited by the degrees of freedom which must be explored in the ligand's conformational space, it is very useful to be able to limit the spatial volume within which docking will be attempted through the preliminary definition of a substrate binding site. In this chapter we will combine a knowledge based calculation (evolutionary residue conservation) and a structure based calculation (accessible volumes) to identify possible substrate access pathways in the Coq6 structure as the pre-requisite step to molecular docking of the substrate. Docking algorithms will explore the conformational space of the ligand in the spatial context of the receptor and calculate some type of energy or energy-like score in order to decide which resulting ligand poses to retain. The best scoring ligand poses are then returned to the user for further evaluation. The quality of the algorithm can be evaluated by its ability to reproduce ligand poses from co-crystallized examples. This indicates that the algorithm's conformational search function can find the proper conformation, and that its scoring function has captured the essential features of ligand-receptor interactions. Optimally, the algorithm will return the crystal pose of the ligand as the top ranked result. In AutoDock Vina, the results are ranked by their energy in kilocalories per mole as computed by a scoring function including terms for hydrogen bonding, hydrophobic interactions, VDW repulsion, and ligand torsions.

This type of confirmation is easy to obtain when reproducing the coordinates of an experimentally solved receptor-ligand pair. However, in the case of Coq6, there is no known experimental structure of the enzyme or of any strictly homologous enzyme-substrate complexes. We will rely much more on the ability of the algorithm to find and rank ligand poses because we have no experimental references. In order to see how much we can rely on the AutoDock Vina ligand pose rankings, we will first perform a preliminary test of the algorithm's ability to find the active site without any extra information.

2.4 Blind docking of 4-HP (polyprenyl length = 0)

The reactive center of the substrate is the aromatic head where the C5 carbon is hydroxylated by Coq6. Therefore our first and simplest test will be to assess the ability of AutoDock Vina to place the aromatic head in the active site of the Coq6 models, without the isoprenyl chain being present to interact with the protein. For simplicity, we used 4-hydroxyphenol (4-HP) as a minimal model for the aromatic head. We present a summary of the results in **Figure 4.2** below.

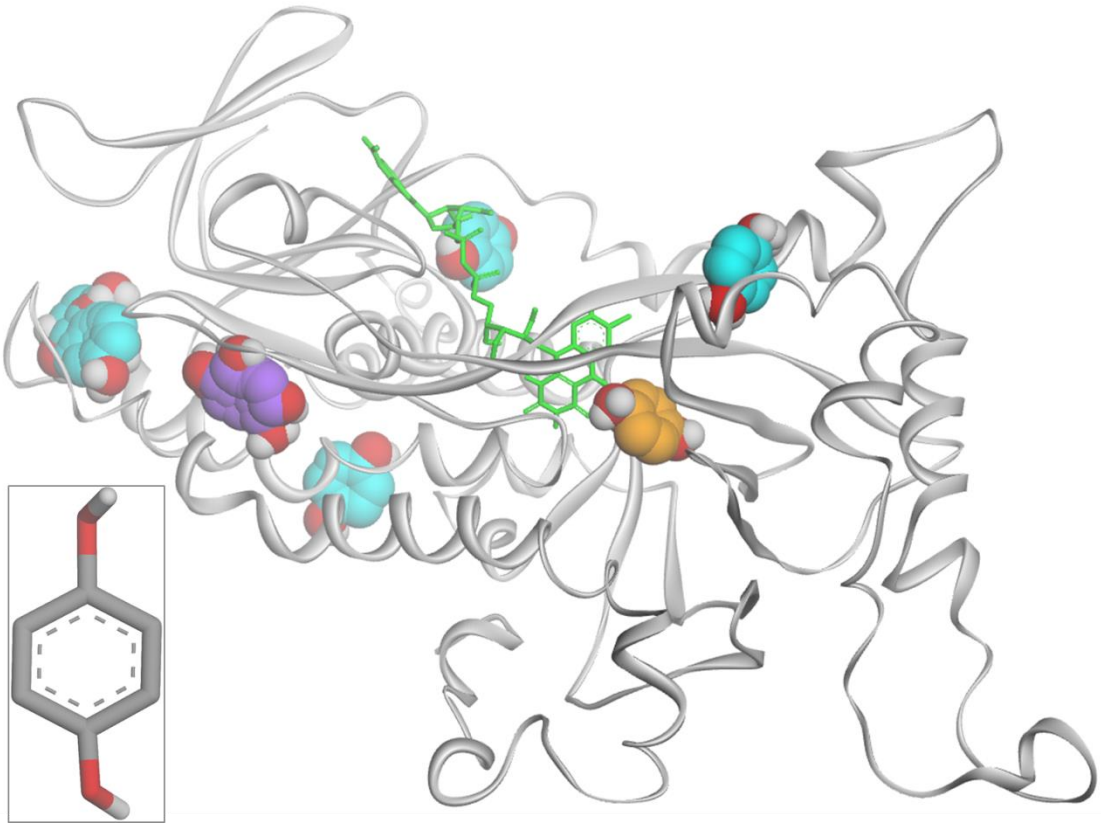

FIG 4.2: *The top 20 ligand poses for blind docking of the aromatic head (space-filling representation) into a Coq6 model structure (shown here is the RATIONAL model) cluster into 6 possible binding sites, only one of which (shown in orange spheres) is catalytically plausible according to its proximity to the FAD isoalloxazine (shown in green stick), despite the fact that the top ranked poses occupy a distal location (purple spheres).*

The top 20 docking results within an energy range of 3 kcal/mol find six major binding sites for the isolated aromatic head. However, only one of these sites is catalytically plausible due to its proximity to the FAD isoalloxazine, and it is populated by ligand poses ranked 11 and 15 out of 20, as represented in orange spheres in **Figure 4.2**. However, 4 of the top 5 ranked poses are in the site shown in purple spheres. This is an important result because it suggests that the docking algorithm is capable of finding the active site, but will not necessarily rank it as the best pose. This is because the experimental coordinates of a bound ligand pose are the result of the *free energy* of the system, which is different from the *energy* computed through techniques likes molecular dynamics, and more coarsely approximated in docking. The difference between the two is why the terms of a docking energy scoring function must be parameterized.[3] However, this parameterization does not capture all of the *free energy* contributions.

For the specific case of the Coq6 system, this is less problematic than it may appear, because we know that the active site is directly in front of the FAD isoalloxazine. Indeed, we can even extract an additional piece of knowledge from one of the template structures: the 1PBE structure of PHBH co-crystallized with FAD and its substrate. This structure gives us a specific distance between the FAD isoalloxazine C4 carbon and the target carbon on the substrate to be hydroxylated. Before proceeding to a more "guided"

docking informed by our tunnel calculations, we will test the ability of the docking program to find the active site when presented with a substrate model bearing a hexaprenyl tail, which is the main distinguishing feature of any Coq6 substrate.

2.5 Blind docking of 3-hexaprenyl-4-hydroxyphenol (polyprenyl tail length = 6)

In this round of docking we added a hexaprenyl tail to the 4-hydroxyphenol head. Once again, we did not specify any specific volume of the receptor as a binding site, allowing the algorithm to attempt docking over the entire Coq6 models. Below we present the results of this round of docking.

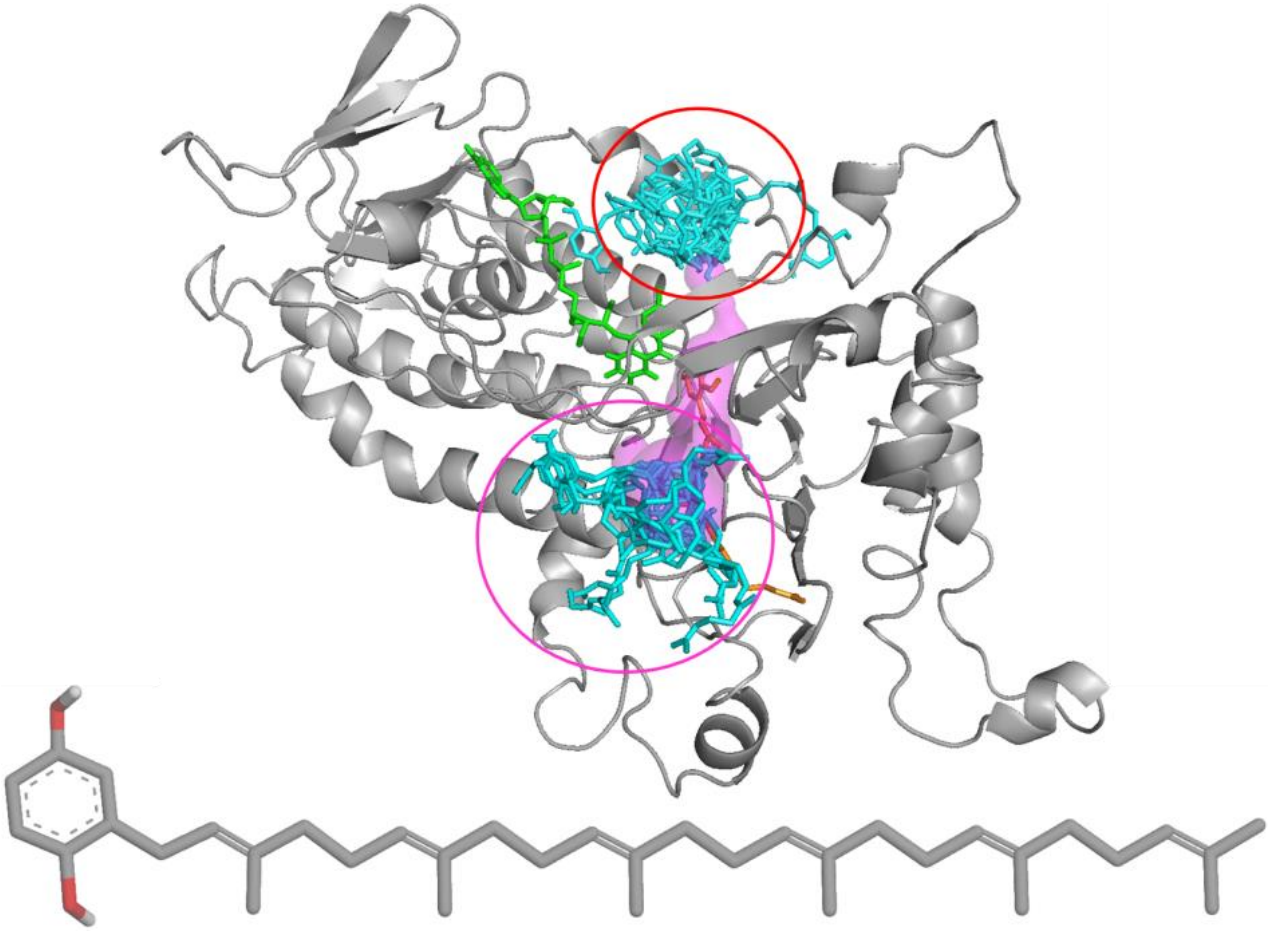

FIG 4.3: *The top 20 ligand poses for blind docking of the model substrate 3-hexaprenyl,4-hydroxyphenol (represented as cyan sticks and enlarged in the inset) into the multi-template Coq6 model. Again, the majority of poses found (cyan sticks) are not near the active site, although one cluster is near the entrance of* re *face tunnel 1, indicated as the purple volume The top ranked poses are colored in purple, and the only catalytically plausible pose is colored in orange. FAD is in green*

The top 20 docking results cluster around three major sites: one is at the entrance of the *re* face tunnel 1, a second one is in a pocket near the *si* loop, and a third one is a single pose in the active site placing the aromatic head in front of the FAD isoalloxazine (shown in orange stick in **Figure 4.3**). This pose places

the aromatic head of the hexaprenylated substrate at the same position as the previous docking of the aromatic head alone (shown in pink). This is highlighted in **Figure 4.4** below.

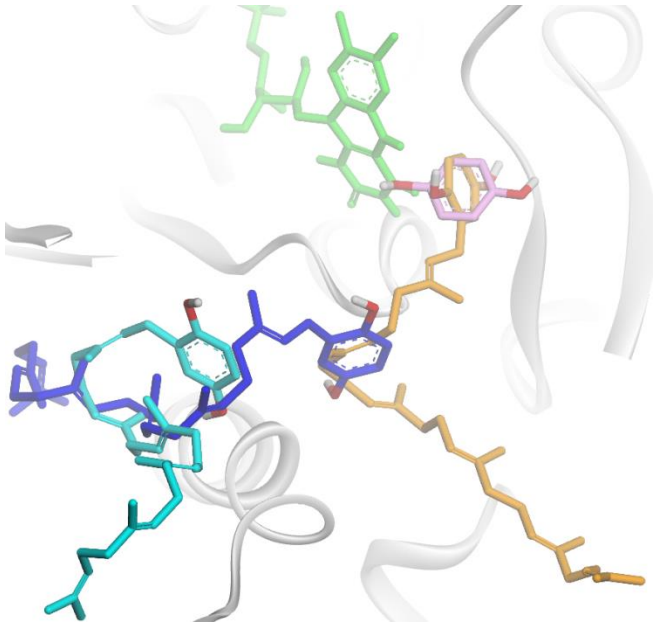

FIG 4.4: *Selected poses for blind docking of the model substrate 3-hexaprenyl,4-hydroxyphenol. Orange: pose rank 4; dark blue: pose rank 9; light blue: pose rank 7; pink: the coordinates of the lone aromatic head from the previous docking run.*

The results also show two other interesting conformations, shown in dark blue and light blue in **Figure 4.4**. These conformations show partial entry of the hexaprenylated substrate to the active site via the *re* face. Despite being catalytically relevant, none of these conformations were ranked as clearly lower in energy than the others. It is interesting to note that in the poses shown in **Figure 4.3**, only the *re* face tunnel entrance is populated by a cluster of substrate models. The other tunnel entrances are apparently less favorable in energy than the inside of the *si* loop. Since the *si*-loop is of highly variable length and composition in the Coq6 family, and not proximal to the active site, we are unsure of the functional significance of this result.

This tells us that for our system the Autodock Vina docking algorithm can find poses that are catalytically plausible, but does not rank them as the best results according to its energy score. Fortunately, we can make knowledge-based selections of likely substrate poses based on proximity of the aromatic head to the FAD isoalloxazine.

2.6 Variations of tail length for computational and experimental approximations

The study of enzyme-substrate interactions is a potential point of interaction between molecular modeling and experimental approaches to the study of the Coq6 system. As described in the introduction, one of the fundamental challenges of studying Coq proteins in general and the Coq6 system in particular is the low solubility of the enzyme and of the substrate. The experimental isolation

and purification of Coq6 has included some modifications to the Coq6 protein, such as the truncation of the N-terminal mitochondrial signal sequence, as well as the addition of a maltose binding protein domain, used to increase solubility and assist crystallization. It is also possible to modify the substrate to increase its solubility, namely through the shortening of the hexaprenyl tail, which is an important step for developing *in vitro* activity assays. Therefore, we decided to examine the effect of isoprenyl tail length on the ability of the aromatic head to attain a catalytic position.

2.7 Site directed docking of 4-HP with tail lengths of 1-6 isoprene units

Encouraged by the ability of molecular docking to place a hexaprenylated substrate model in a catalytically plausible position, we decided to explore docking of substrate models with shorter polyprenyl tails, since solubility is inversely related to tail length. We decided to systematically explore substrates with different polyprenyl tail lengths, ranging from 1 to 6 isoprene units. This exploration will allow us to determine the optimal polyprenyl tail length for our *in silico* model substrate, as well as for experimental substrate models.

The results from the previous step of docking indicate that while the Vina docking algorithm can find catalytically plausible poses, these are a small minority of the poses found without restricting the search volume on the receptor. Therefore, in order to find more relevant ligand conformations, we decided to restrict the docking volumes to accessible volumes (which we will describe in further detail in Section 3.3). An example of a docking box is shown below in **Figure 4.5**. This box measures 45 x 35 x 27 Å, enclosing a total volume of 45 525 $Å^3$.

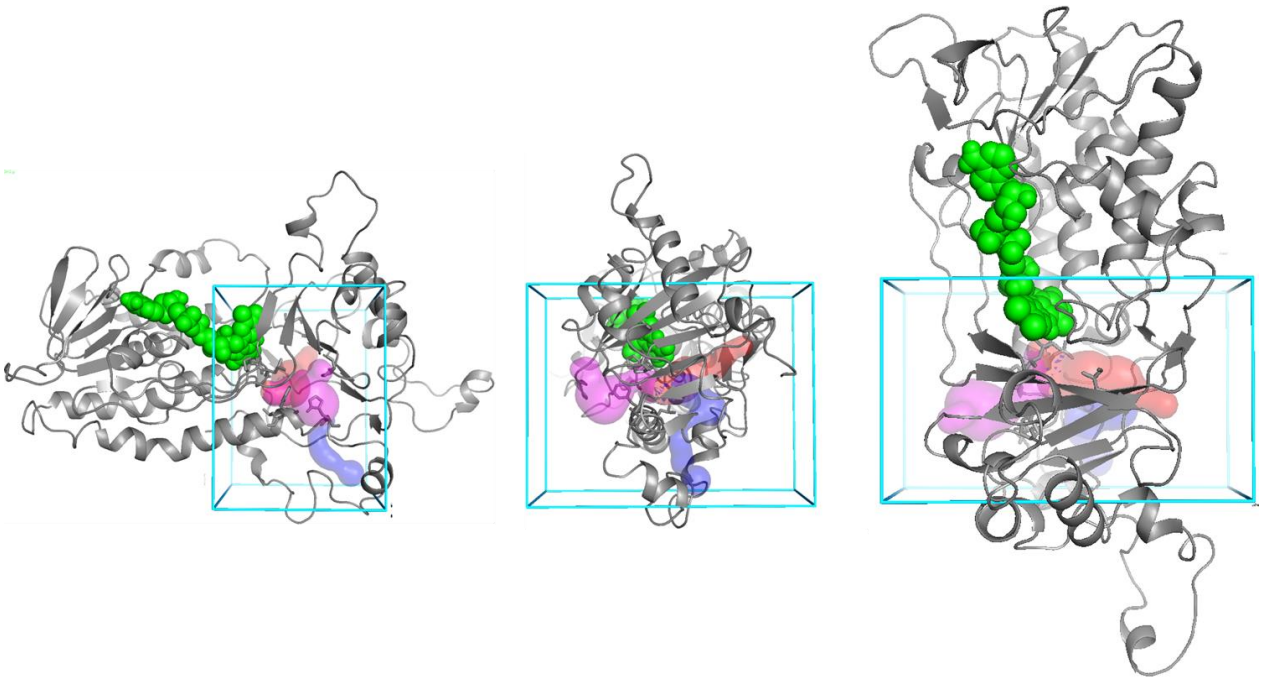

FIG 4.5: *The box defined to restrict docking attempts to the tunnel system identified in the Coq6 models. Shown here: the RATIONAL Coq6 model. FAD is shown in green sphere. Purple volume: re face tunnel 1. Blue volume: re face tunnel 2. Red volume: the si face tunnel.*

We proceeded to dock model substrates with varying tail lengths into a conformation selected from MD of the RATIONAL model to assess the resulting positions for the aromatic head. Here we present only the results from docking into the RATIONAL model because it has the best formed tunnel system, as will be explained in the following sections. These first results are resumed in **Figure 4.6** below. Model substrates with one isoprene unit are hereafter referred to as Q1; substrate models with two isoprene units are called Q2, and so on to Q6, the *in vivo* length of the substrate in *S. cerevisiae*.

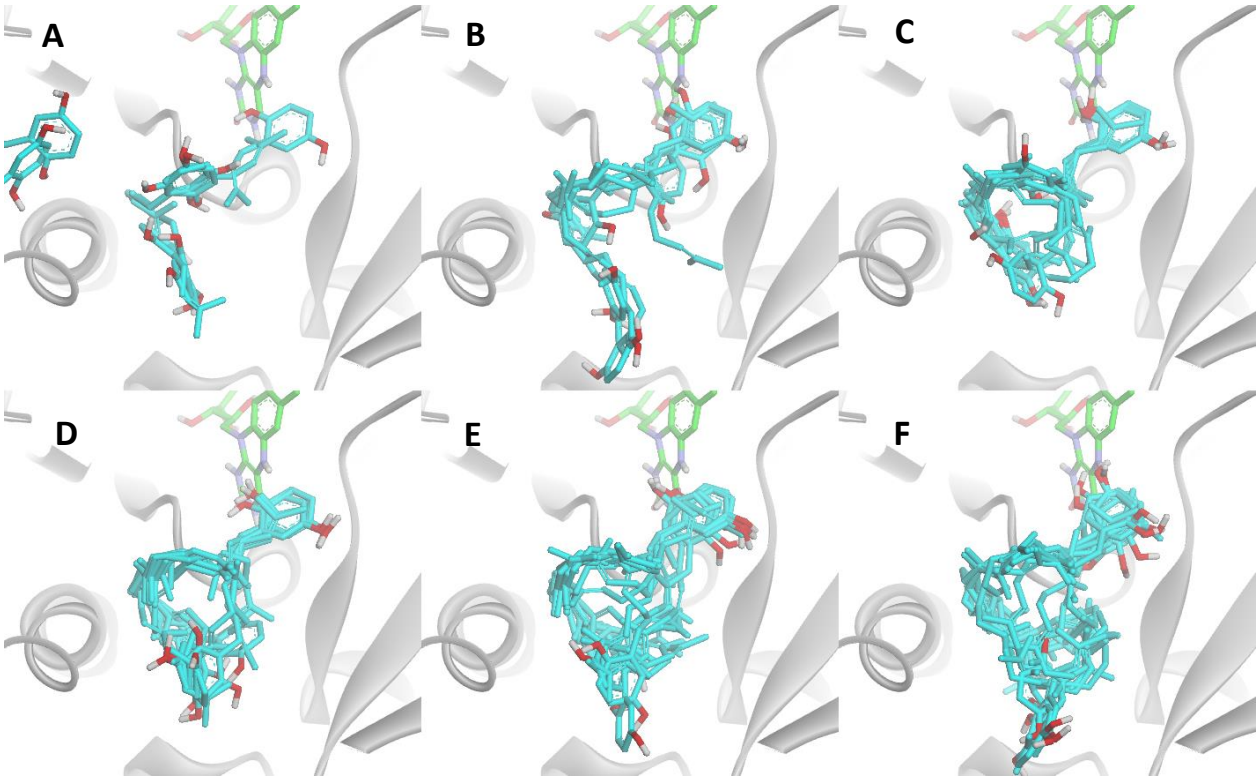

FIG 4.6: *Docking of substrate models with progressively longer polyprenyl tails. A) Q1, B) Q2, C) Q3, D) Q4, E) Q5, F) Q6. Only the top ten poses for each model of tail length are shown and superposed (cyan sticks). The FAD isoalloxazine C4 atom, which will bear the reactive peroxo group, is shown as a pink sphere. The enzyme model in this figure is the RATIONAL Coq6 model based on 2X3N and 4N9X.*

We can observe an interesting trend: for shorter tail lengths there is more variability in the position of the aromatic head, especially visible in **Figure 4.6**, including many poses where it is nowhere near the FAD isoalloxazine. This suggests that the active site volume in front of the FAD is relatively large, much larger than necessary for just the aromatic head. The size of the active site volume in the Coq6 models is mainly a result of the Coq6 sequence, as much of the active site backbone geometry is inherited from the 2X3N template, which does not show such a large tunnel system. The variability in position of the 4-hydroxyphenol with one isoprene unit suggests that in real life, short tailed substrate models may be able to enter the active site, but not maintain a catalytically plausible pose.

Increasing the polyprenyl tail length to Q2 results in a tighter distribution of docked poses, although there are still several poses where the aromatic head is distant from the FAD. Q3 shows a tighter

clustering of the poses closer to the FAD. This trend continues with Q4 through to Q6, where the aromatic ring becomes more and more consistently localized in front of the FAD and the entire polyprenyl tail is able to fit inside the active site volume. Since increasing the length of the isoprenyl chain to 6 units results in better and more consistent positioning of the substrate's aromatic head, we will proceed with a substrate model with a full length, six unit polyprenyl tail. These results also suggest that the minimal tail length is four isoprene units, offering a potential balance between solubility and catalytic reactivity.

2.8 Docking survey conclusions

This preliminary docking survey demonstrates the importance of *a priori* knowledge of the active site for the interpretation of docking results. We have also briefly introduced two criteria for selecting appropriate receptor conformations from molecular dynamics trajectories: the location of the active site, and the calculation of accessible volumes forming a contiguous and traversable pathway for the substrate from the exterior of the enzyme to this buried active site.

In the case where the substrate is a relatively small molecule that does not extend far beyond the atom being operated on, the active site is also the complete definition of the substrate binding region. However, the case of Coq6 is different. Its substrate, a prenylated Q biosynthesis intermediate, is a relatively large substrate molecule in the sense that most of its mass and volume is represented by the polyprenyl tail, which is not the region of the molecule undergoing catalysis and does not participate actively in the process. Yet, because the polyprenyl chain is covalently attached to the aromatic head, it is clear that the tail will have to make some contacts with the protein if the aromatic head is to reach the catalytic center of the enzyme. This position, which can be defined as the location of the C4a atom of the FAD isoalloxazine (which bears the reactive peroxo group during catalysis) is buried near the center of the protein nearly 14 Å away from the protein surface, as shown in **Figure 4.7** below.

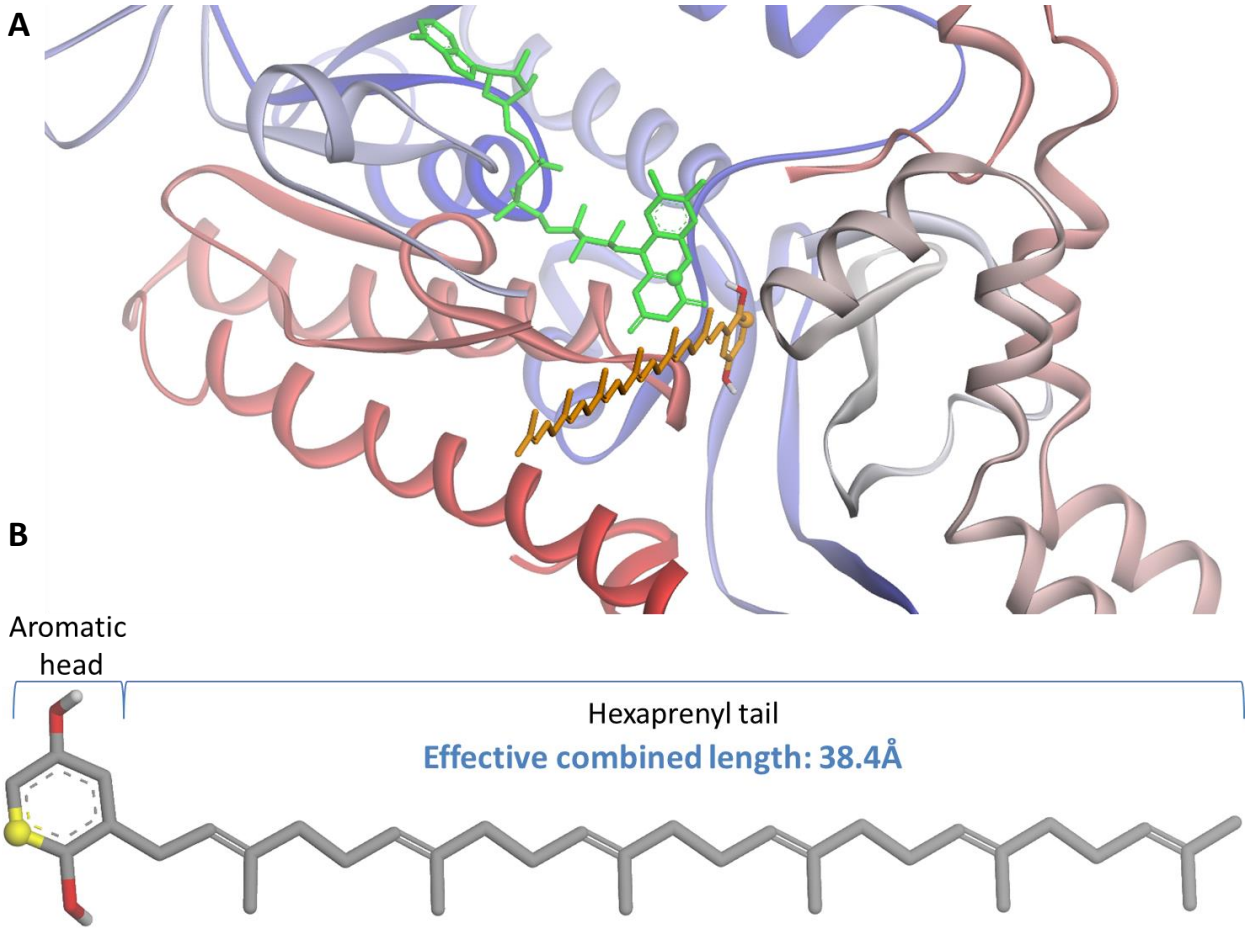

FIG 4.7: *A) The active site of the Coq6 RATIONAL model with FAD in green stick. A hexaprenylated model substrate 3-hexaprenyl-4-hydroxyphenol is shown in orange stick, manually positioned in an extended conformation to place the aromatic head in the active site, proximal to the FAD isoalloxazine. The FAD C4a atom is in green sphere; the substrate C5 carbon is in orange sphere. B) A closer view of the model substrate, C5 carbon in yellow sphere. C) The active site of PHBH structure 1PBE, co-crystallized with FAD and the substrate, para-hydroxybenzoate*

Therefore, we analyzed our models for the presence of a physical pathway for the substrate's aromatic head to reach the active site while bearing the polyprenyl substituent. This analysis is based on the criteria of evolutionary residue conservation, molecular geometry, and hydrophobicity. This next section of this chapter describes the analysis that will identify the residues of the active site and any associated substrate binding region of the enzyme, which are required to define a specific region for substrate docking.

3. Enzyme model analysis

3.1 Active site identification

The 3D position of the Coq6 active site is trivially easy to define: it is necessarily located immediately in front of the FAD isoalloxazine (as shown in **Figure 4.7**), because this is where the reactive peroxo group is added to flavin during the reaction cycle. This is confirmed through the many crystal structures of PHBH which have been crystallized with substrates, substrate analogs, or products.[4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] Indeed, these enzyme-substrate complexes can give us a specific interatomic distance between the FAD's C4a atom and the target carbon of the substrate. However, inspection of this region among the three remaining Coq6 models makes it obvious that the enzyme must have some way of permitting passage of the polyprenyl tail if the aromatic head is to attain this position. Therefore we will analyze another aspect of the Coq6 structure: the conservation of residues among proteins of similar sequence.

3.2 Evolutionary residue conservation

While the general location of the active site in the Coq6 models is trivial, the identification of a possible substrate channel is not. Specific binding of a given substrate imposes geometrical and evolutionary constraints upon the enzymes which operate on them. This is likely to be especially true for Coq6 substrates, which have a large hydrophobic tail. This suggests that Coq6 may have a structural adaptation to such a substrate which should be detectable in its molecular structure.

The most reliable experimental data we have on the Coq6 molecular structure is its amino acid sequence. This gives us the basis for an analysis of evolutionary residue conservation through the multiple sequence alignment of similar sequences, which we performed with the ConSurf[15] method. This method uses the MSA of close Coq6-family sequences (introduced in Chapter 3, and presented in full in Annex 2) to compute a conservation score for every residue, which are then mapped onto our homology models in order to cross-reference them with protein model geometry analyses.

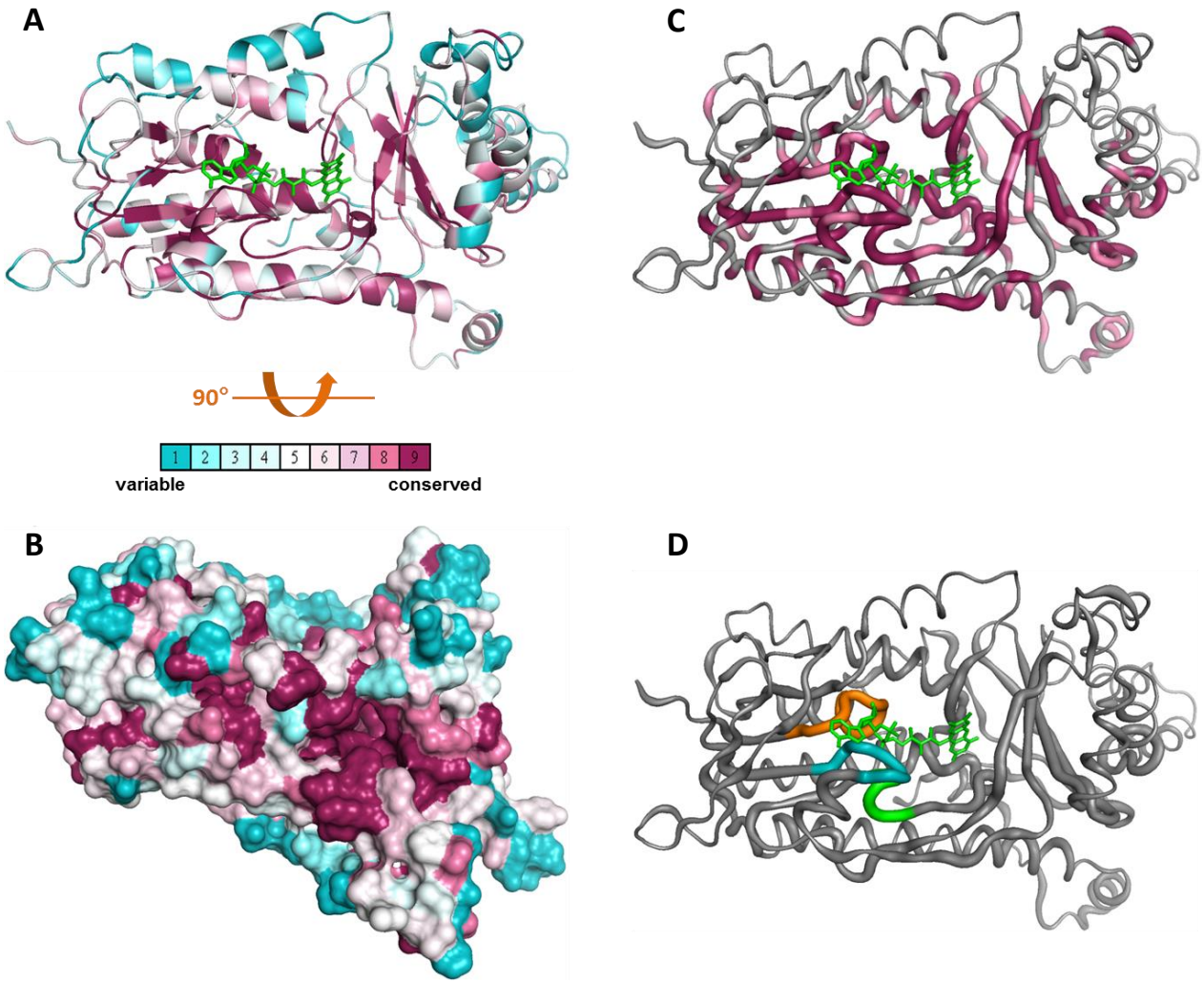

FIG 4.8: *Evolutionary residue conservation as calculated by the ConSurf method. The scores run from 1 (most variable) to 9 (most conserved) and are assigned a color on the color scale, which are projected onto a 3D model (here, the RATIONAL model. A) Ribbon view from the "top" of the FAD (in green stick) showing residue conservation in the FAD binding site. Also visible is residue conservation in the large beta sheet immediately in front of the isoalloxazine. B) Surface rendering seen from the re face of the enzyme, showing strong conservation of spatially proximal residues around the large beta sheet forming a surface depression. C) Same view as A, with all residue colors set to gray except for the most conserved (score of 9) residues, in purple. Backbone width is proportional to residue conservation. D) Same view as A, with all residue colors set to gray except for the ADP binding motif (orange), the NAD(P)H binding motif (blue) and the ribityl binding motif (green).*

The projection of the residue conservation scores onto the Coq6 model confirms the ability of the ConSurf method to identify sequence motifs that correspond to known functional structures of this global fold (as shown in **Figure 4.8D**):

- the ADP binding motif (GxGxxG), which is IVGGGPAGL in the Coq6 sequence
- the NAD(P)H binding motif (GxDGxxx), which is GAGFNS in the Coq6 sequence
- the ribityl binding motif (GDAxH), which is GDAAH in the Coq6 sequence

The three conserved FAD binding motifs shown in **Figure 4.8** have residue conservation scores of 9, the maximum on the ConSurf scale. The ConSurf analysis also detects two other regions of the protein as being highly conserved: the large beta sheet, typically involved in substrate binding in this class of proteins, contiguous with a patch of conserved residues forming a depression on the surface of the protein. This recapitulates a similar structural feature found in the structurally and functionally homologous enzyme, PDB entry 4N9X, which is a Q biosynthesis monooxygenase.

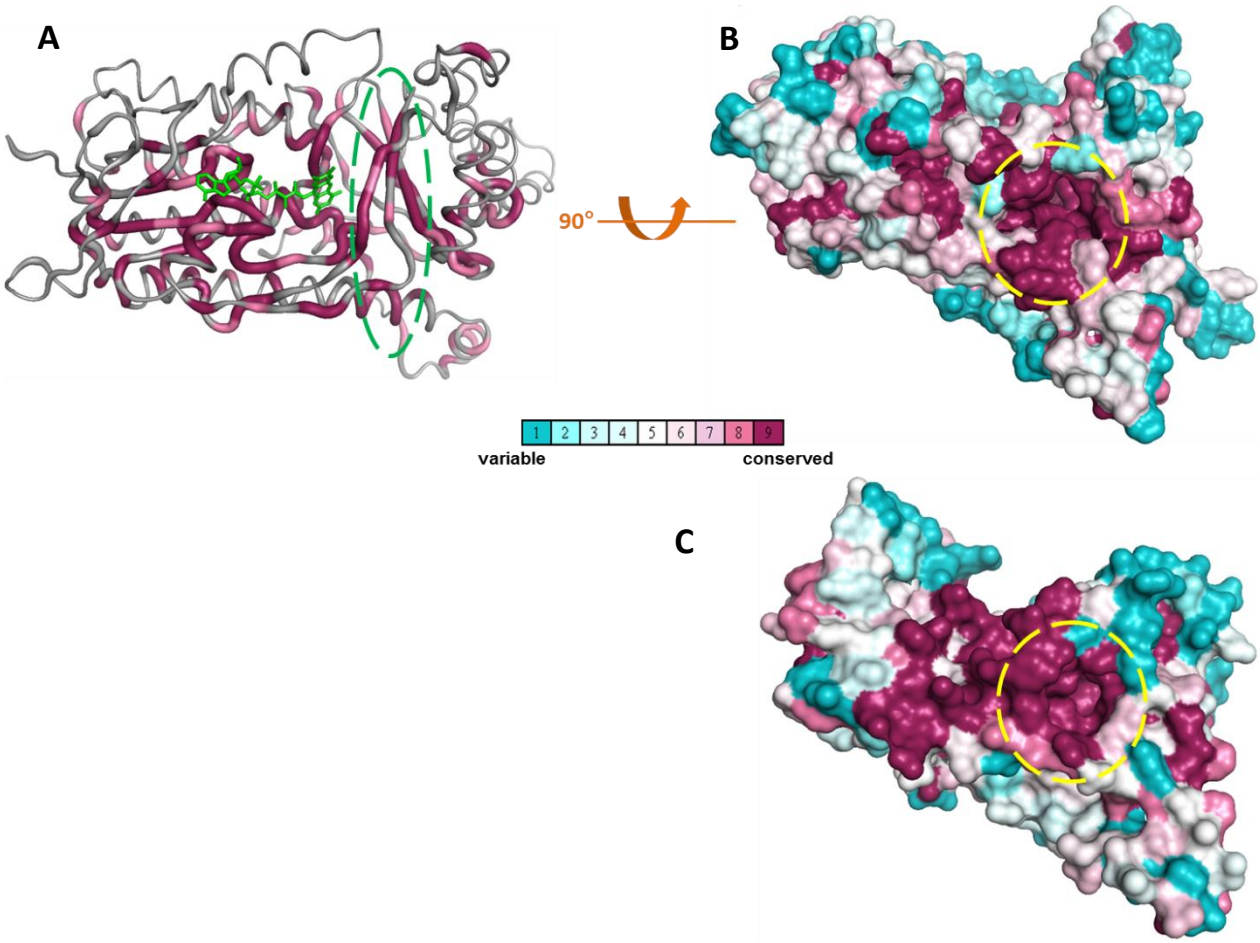

FIG 4.9: *Evolutionary residue conservation (projected here on the RATIONAL model) highlights two structural features of the Coq6 family of proteins: A) the large beta sheet of the beta sheet domain (highlighted in the green oval), and B) a surface depression proximal to the active site (highlighted in the yellow oval), which is recapitulated from C) the crystal structure of one of the templates, 4N9X, a Q biosynthesis monooxygenase (also highlighted in the yellow oval).*

The beta sheet domain is likely to be involved in substrate interactions, as is the case for most examples of this type of enzyme. Since the MSA used to compute the residue conservation scores was limited to the Coq6-family enzymes, we can infer that all the enzymes represented in the MSA results all have to bind a hexaprenylated Q biosynthesis intermediate. The MSA results for the residue conservation analysis performed on both the 4N9X template as well as a Coq6 model reveal a common feature: a surface depression formed by a set of highly conserved residues (circled in dashed lines in **Figure 4.9**).

Inspection of these regions reveals that they are spatially contiguous with each other as well as the position of the active site. This is suggestive of a functionally important region which is contiguous with the active site and the protein surface.

3.3 Accessible volume calculation: CAVER[16]

While residue conservation is a strong clue of functional importance, a hypothetical substrate access channel should also be detectable through calculations of protein structure geometry. This calculation, performed with CAVER[16], identified a channel leading from the active site to the protein surface which is lined by a set of residues identified as strongly conserved by the ConSurf analysis.

The tunnel system computed with CAVER reveals an active site volume contiguous with but distinct from tunnel volumes reaching the surface. The geometry of both of these accessible volumes is significant for characterizing conformations of the Coq6 homology models for use in subsequence substrate docking calculations, which reveal that the polyprenyl tail occupies an access channel and the aromatic head is bound in the active site volume in front of the FAD. Accessible volumes were computed for the three models and reveal **three distinct types of tunnels** which converge to the active site. Two of these tunnels (*re* face tunnels 1 and 2), exit the enzyme via its *re*-face, while the other exits via the *si*-face (the *si* and *re* nomenclature is derived from the *si* and *re* face of the FAD isoalloxazine). For each of the three models, the three types of tunnels were considered as possible substrate access channels to be tested through substrate docking. However, the CAVER calculations also reveal many other tunnels, many of which do not reach the active site and are much too small to admit passage of the substrate. Many of these tunnels, due to their small sizes, collapse below the detectable size limit over the course of the simulation. We term them “ghost” tunnels.

This is resumed in the following figures (**Figures 4.10 – 4.12**), which show for each pre-dynamics homology model two views of the accessible volumes computed by CAVER. The first view, presented in Panel A of these figures, is of the complete set of tunnels, colored in beige. Of these tunnels, we can identify the three main tunnel types. The *re* face 1 tunnel is colored in purple, the *re* face 2 tunnel is colored in blue, and the *si* face tunnels colored in red. The beige tunnels of the complete set are usually of very small diameter, have a complex path through the protein model structures, and do not lead to the active site. While we do not explicitly analyze these voids in the same detail as the three main tunnels identified, they often describe accessible volumes directly proximal to the tunnels of interest and the active site itself. This will become apparent in the presentation of the model substrate docking results in Section 7. To briefly resume, the volume of space designated for docking is defined as a rectangular prism, whereas the tunnels have irregular volumes defined as a twisting path of overlapping spheres. This means that even a minimal bounding box for any given tunnel of interest will necessarily contain accessible volumes attributed to “ghost” tunnels. The principal result is that even in docking calculations intended to test one tunnel at a time, it is not possible to prevent the model substrate from occupying some accessible volume in the other tunnels, which is an effect visible in some of the results.

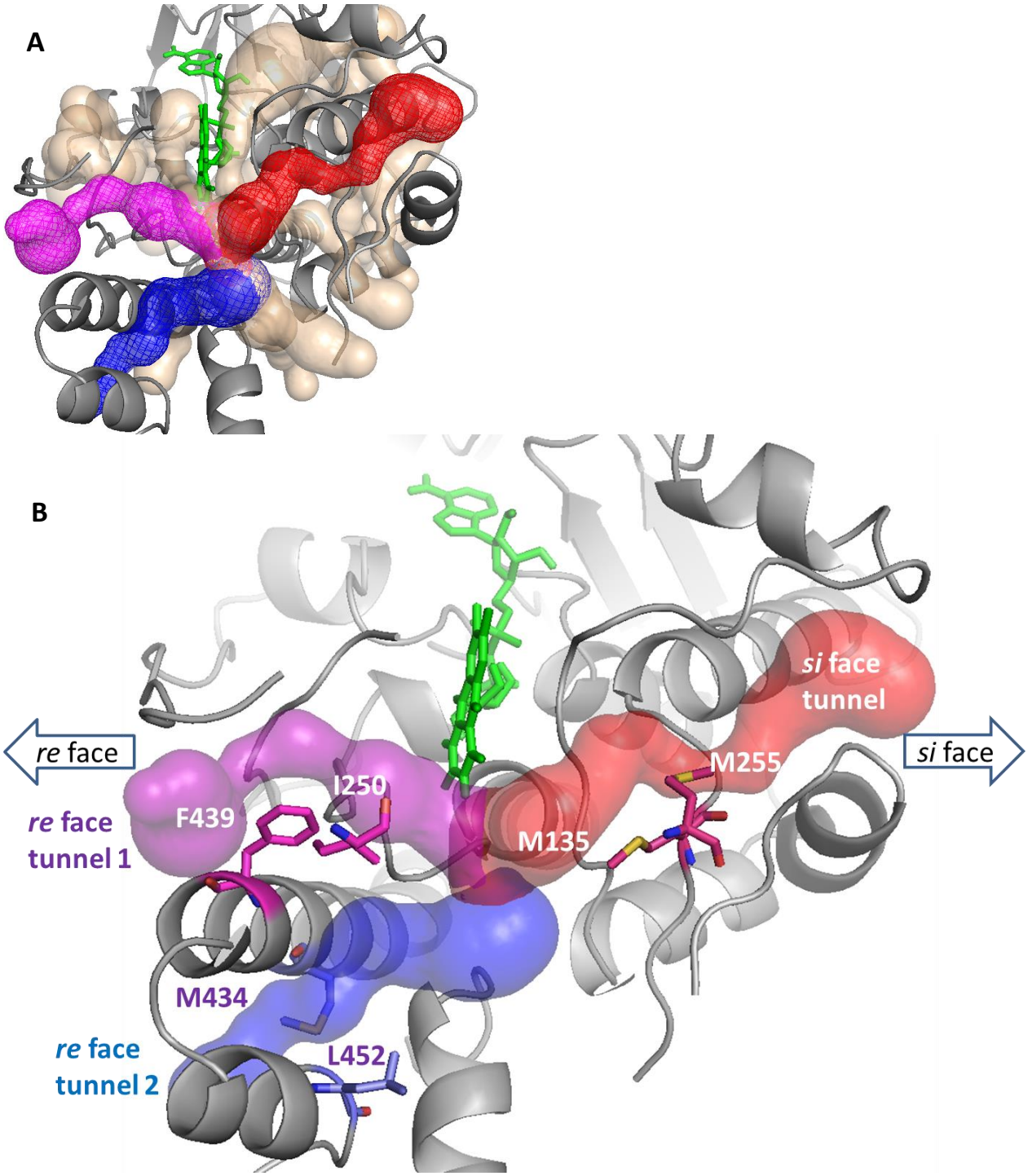

FIG 4.10: *Volume rendering of the channel system in the I-TASSER Coq6 homology model. A) The complete results of the accessible volume calculation by CAVER, showing many minor tunnels that are either too small to admit passage of the substrate or do not converge to the active site. These beige "ghost" tunnels and volumes were manually identified and not considered for substrate docking. B) The three main tunnel types which converge to the active site: re face tunnel 1 (purple), re face tunnel 2 (blue) si face tunnel showing bottleneck residues as sticks.*

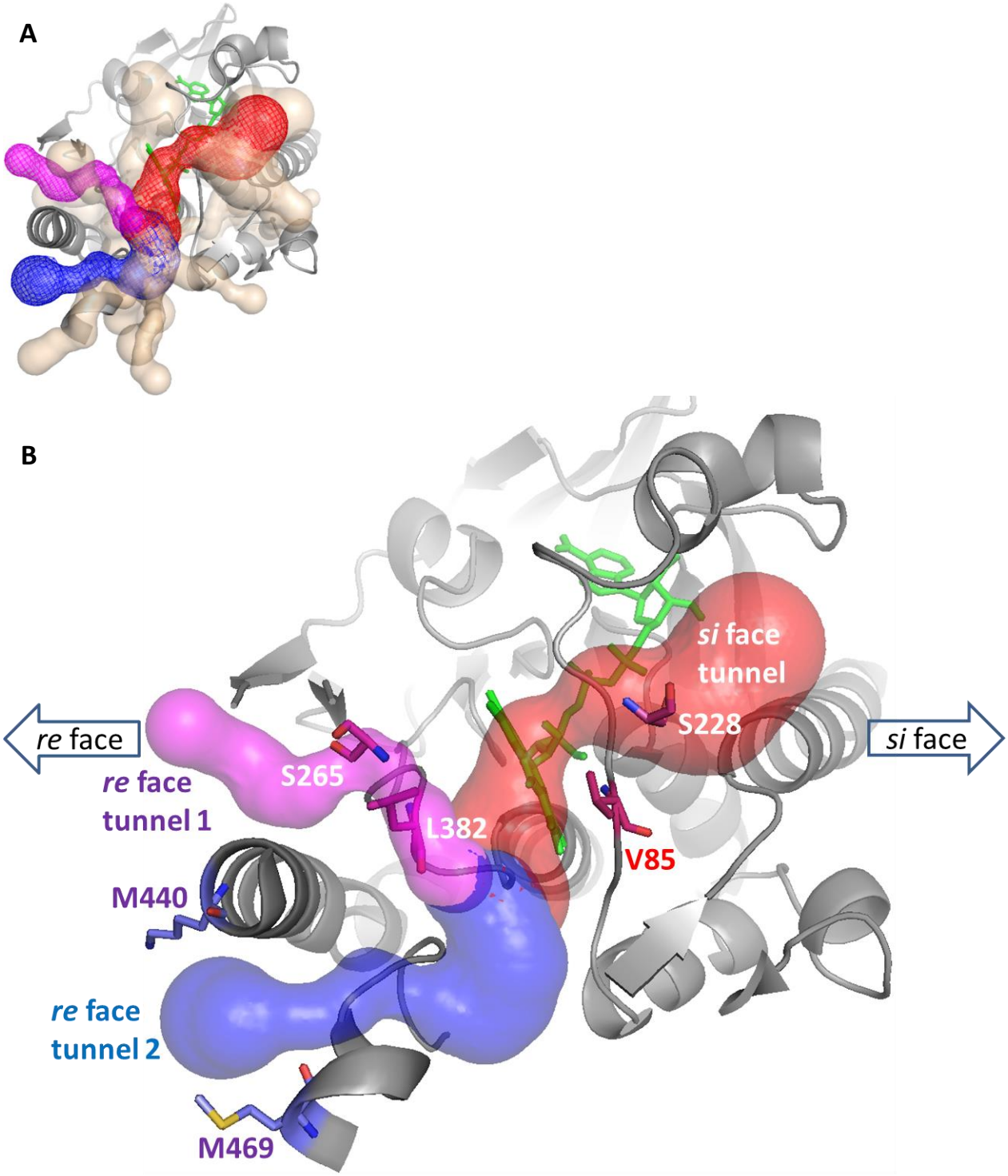

FIG 4.11: *Volume rendering of the channel system in the ROBETTA Coq6 homology model. A) The complete results of the accessible volume calculation by CAVER, showing many minor tunnels that are either too small to admit passage of the substrate or do not converge to the active site. These beige "ghost" tunnels and volumes were manually identified and not considered for substrate docking. B) The three main tunnel types which converge to the active site: re face tunnel 1 (purple), re face tunnel 2 (blue) si face tunnel showing bottleneck residues as sticks..*

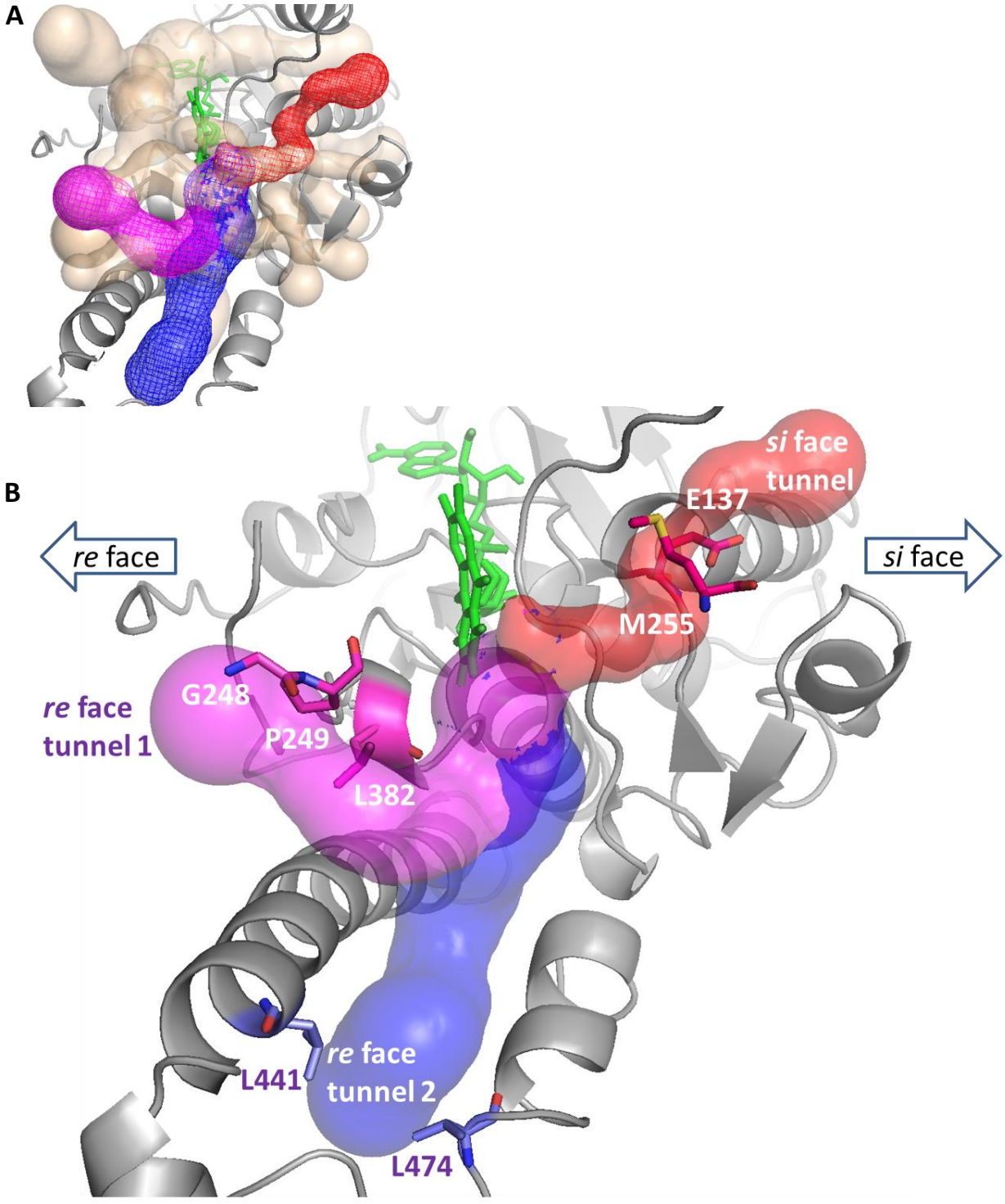

FIG 4.12: *Volume rendering of the channel system in the RATIONAL Coq6 homology model. A) The complete results of the accessible volume calculation by CAVER, showing many minor tunnels that are either too small to admit passage of the substrate or do not converge to the active site. These beige "ghost" tunnels and volumes were manually identified and not considered for substrate docking. B) The three main tunnel types which converge to the active site: re face tunnel 1 (purple), re face tunnel 2 (blue) si face tunnel showing bottleneck residues as sticks..*

The putative function of these tunnels is to permit passage of the substrate's aromatic head from the exterior of the protein to the buried active site. Therefore we are particularly interested in their diameters, as their diameters will be the limiting factor in allowing passage of the substrate to the active site. The tunnel diameters will be measured at their narrowest points, since these points will be the limiting factors for substrate passage in each tunnel. The bottleneck residues of each of the tunnels are shown in **Figures 4.10 – 4.12**. We will develop the calculation of the effective diameter of each tunnel over the course of the molecular dynamics simulations in the following section.

4. Molecular dynamics simulations: effective diameter of the tunnels and substrate

The CAVER method has been essential for the calculation of accessible volumes, some of which we suspect may be substrate access channels.. Before proceeding to a more detailed analysis of accessible volumes we will first describe our choice of a specific substrate molecule, which was rationally selected on the basis of both experimental evidence and theoretical prediction.

4.1 Substrate model selection

Previous work in the field proposed 3-hexaprenyl-4-hydroxybenzoate (4-HB6) as the substrate of Coq6 on the basis that the decarboxylation reaction of the eukaryotic Q biosynthesis pathway *could* occur *after* the C5-hydroxylation step catalyzed by Coq6.[17] However, more recent work in the field by our group showed that cells lacking a functional Coq6 enzyme accumulate a different species: 3-hexaprenyl-4-hydroxyphenol (4-HP6).[23] This indicates that in the *in vivo S. cerevisiae* system, the C1-decarboxylation and hydroxylation can occur independently of the C5-hydroxylation catalyzed by Coq6. However, since there are no *in vitro* assays for the C1-decarboxylase, the C1-hydroxylase (neither of which has been identified), or the C5-hydroxylase (Coq6), there is no unambiguous experimental evidence to identify either 4-HB6 or 4-HP6 as the substrate of Coq6.[18] Analysis of the active site of PHBH, an extensively characterized enzyme with structural (global fold) and functional (aromatic hydroxylation) homology to Coq6 (both are Class A flavoprotein monooxygenases) reveals that the carboxyl group of the PHBH substrate (para-hydroxybenzoate, or 1-carboxyl-4-hydroxybenzene) is hydrogen bonded to the guanidinium group of R214. However, the Coq6 active site has no residue homologous to PHBH R214, suggesting that a carboxyl group may not be present on the aromatic head of the Coq6 substrate. Together, these observations suggested that 4-HP6 was a more likely substrate for Coq6. While we used both 4-HB6 and 4-HP6 in our docking tests, we found a greater diversity of poses for the carboxylated head of 4-HB6 than for the di-hydroxylated head of 4-HP6 when docking into the same frames extracted from MD. This was because of the lack of a Coq6 equivalent to the PHBH R214 residue. In contrast, the two hydroxyl groups of 4-HP6 were found to form hydrogen bonds with conserved Coq6 active site residues, limiting the diversity of favorable positions for the aromatic head and placing it consistently near the FAD isoalloxazine C4a atom. Therefore, we will only present the substrate docking investigation using 4-HP6.

4.2 Tunnel diameter estimation

The CAVER method and algorithm has been essential for the calculation of accessible volumes (tunnels) and identifying bottleneck residues in each of these tunnels. However, the CAVER algorithm has some drawbacks for our application to Coq6 MD trajectories. They arise mainly from the use of a probe of fixed radius to calculate accessible volumes.

First, a probe of fixed radius (0.9Å) introduces a subtle but important assumption in the interpretation of the results: it implies that the tunnel has a circular cross section, when in fact it does not. It also implies that the substrate has a circular cross section, when in fact it does not. The substrate has different dimensions on different axes of its cross section, as illustrated in **Figure 13** below.

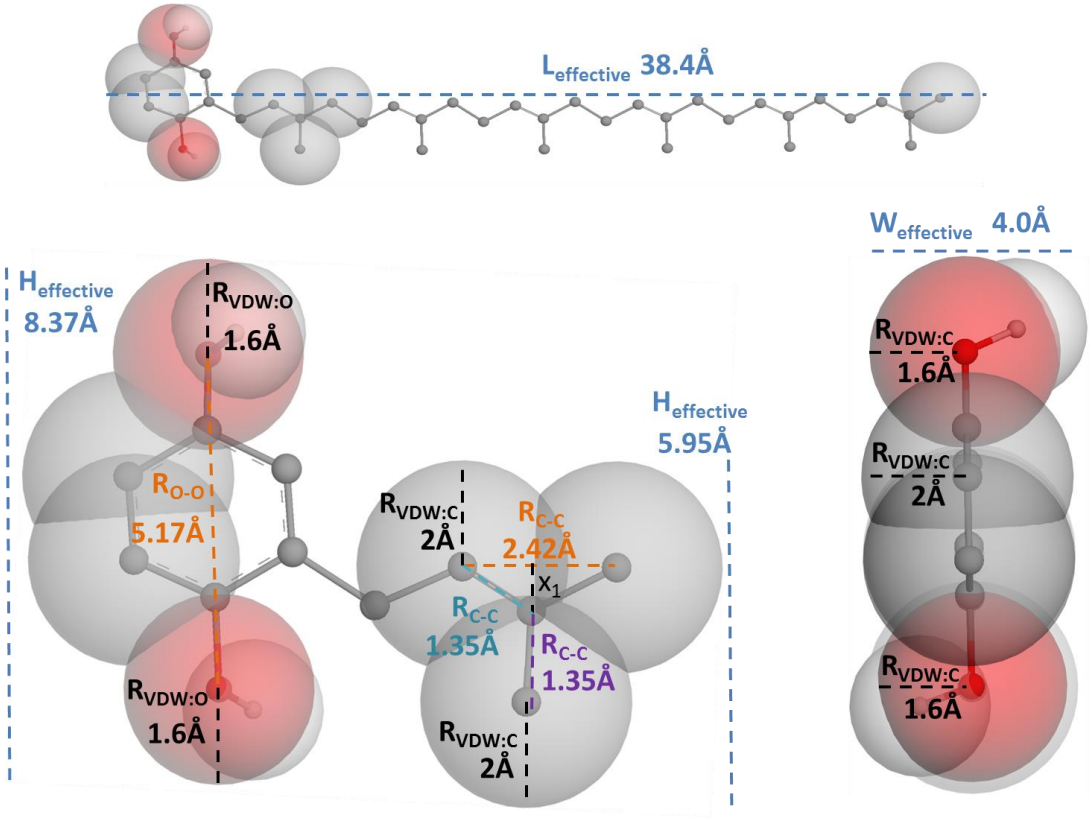

FIG 4.13: *The effective dimensions of a model substrate (3-hexaprenyl-1,4-dihydroxyphenol) on different axes. While tunnel choke points have been characterized by a single diameter measurement, the substrate can present several possible effective diameters depending upon the relative orientation between substrate and tunnel. VDW radii are those used by AutoDock Vina.*

It also means that if the tunnel diameter shrinks below the probe diameter, the tunnel will not be detected. This has the effect of artificially "quantizing" the detection of the tunnel to the diameter of the probe, with its underlying assumption of circular cross section, which is not true for the tunnel or the substrate.

In practice, this means that over the course of the molecular dynamics trajectory, a tunnel, or entire systems of tunnels can disappear abruptly from one frame to another. The accessible volume itself may not have disappeared, but it will no longer be detected by the algorithm because the volume will not be calculated as accessible to a probe of a given radius.

One solution is to specify probes of smaller sizes in order to track the diameter of the tunnel. However, this will lead to the detection of a much larger number of very small voids, most of which do not correspond to functional features of the enzyme. This slows down the calculation, since many more

tunnels must be computed. While the tunnels are automatically clustered, they must be manually inspected and the smaller tunnels must be discarded, further slowing the analysis. This solution is essentially a compromise on the "spatial resolution" of the CAVER algorithm.

Another solution is to perform the volume calculation at larger time intervals in the trajectory. However, this has the drawback of potentially missing enzyme conformations that are suitable for substrate docking, despite their presence in the trajectory. This solution essentially is a compromise on the "temporal resolution" of the CAVER algorithm.

We found that neither solution was satisfactory. Specifying a smaller probe could theoretically allow finer spatial resolution of the tunnels, but in practice it is prohibitively slow for our trajectory analysis needs (we will need to perform the analysis 9 times, because there are three tunnels in each of the three models). Specifying a larger time interval for trajectory sampling has the drawback of potentially missing substrate binding conformations that have in fact been computed. Further, specifying large time intervals for CAVER calculations also makes assumptions about the timescales of tunnel diameter fluctuations, of which we know nothing *a priori*.

We want to **characterize the bottleneck diameters of these tunnels** at the highest spatial and temporal resolution available to us through our simulations. Our MD trajectories were computed with all atoms (including non-polar hydrogens) at a timestep of 2fs. Since we do not have any knowledge of the timescale of interest for tunnel dynamics, we would like to measure the diameter of the tunnel as continuously as possible over the entire trajectory. Therefore, we would like to have a measurement for every frame in the MD trajectories, which consist of 10 000 frames covering 20 nanoseconds computed at a timestep of 2 femtoseconds. We also need this calculation to be rapid. Therefore we will develop an alternative method for measuring the effective diameter of the tunnel bottlenecks.

The Visual Molecular Dynamics (VMD) software package[19] gives us a simple way of implementing this: the direct measurement of distances between selected pairs of atoms, which is extremely rapid to calculate. There are two caveats to using this method. First, the manual atom selection must correspond to the geometry of the tunnel. Second, since the distances measured are between atomic centers, we must add a correction for the van der Waals radii of the atoms of the bottleneck residues.

4.3 Atom selection

We have seen in **Figures 4.10 – 4.12** that each tunnel contains a point of minimum diameter, called a bottleneck, which will be the limiting factor in permitting passage of the substrate's aromatic head to the active site. We can refine this definition to atomic resolution by selecting an atom from each residue which protrudes the most into the tunnel lumen. Having defined this pair of atoms (one from each bottleneck residue), we can then measure the center-to-center distance between them. Then we must add a correction factor to the measured value to reflect the physical volume of the atom, which will make the effective diameter of the tunnel smaller than the measured distance. In order to illustrate this geometry and then develop the effective diameter calculation which follows, we will present a general case of this measurement taken from the *re* face tunnel 1 of the RATIONAL Coq6 model.

An important point to make at this juncture is that while CAVER can identify residue which form the bottlenecks of each tunnel, it does not specify which atoms contribute most to tunnel occlusion. Since CAVER's method calculates the tunnel diameter (including at the bottleneck) with a probe, it does not

consistently select a particular atom from each residue as a bottleneck atom over the course of the trajectory. There is good reason for this, as we illustrate in **Figure 14** below.

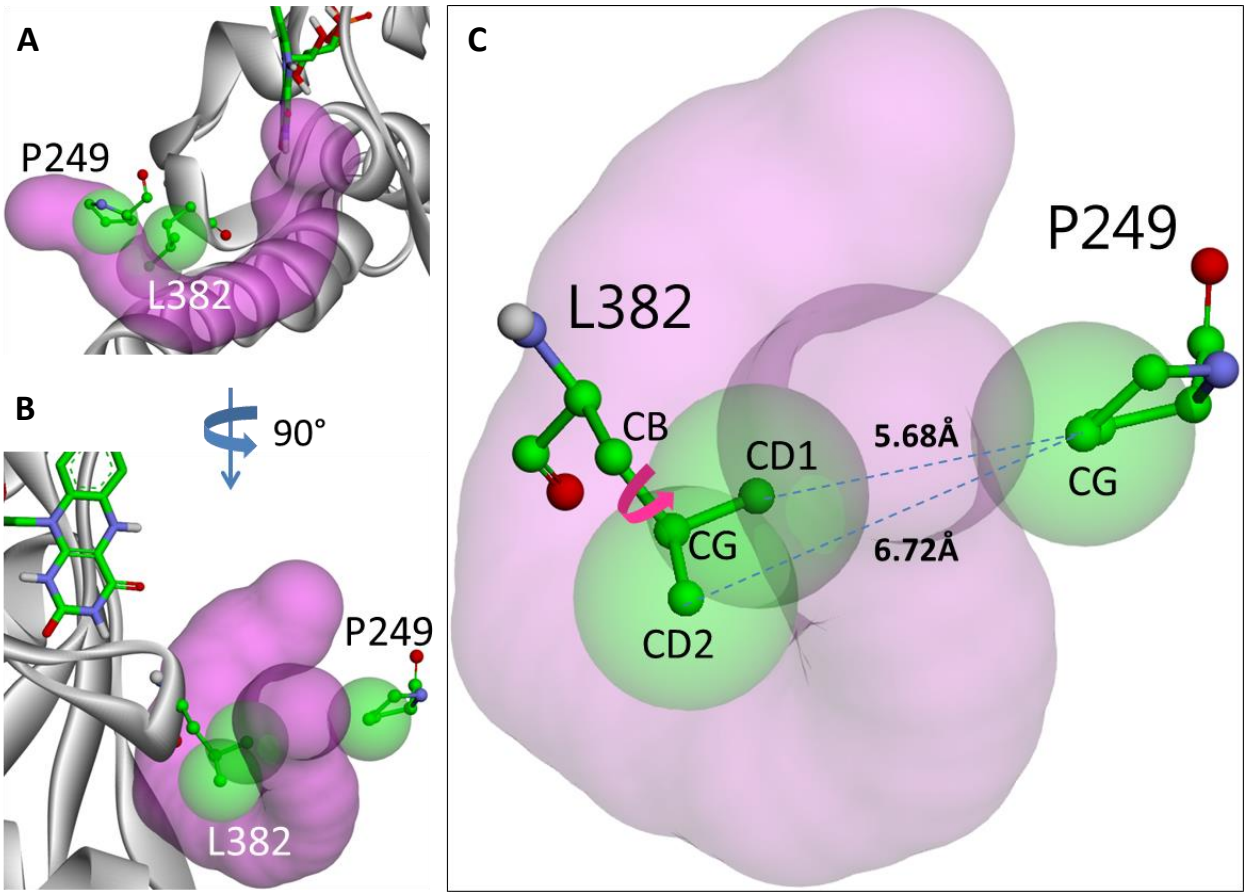

FIG 4.14: *Bottleneck residue rotational degeneracy possible between the CD1 and CD2 atoms of the L382 sidechain, one of the bottleneck residues in the re face tunnel 1 as identified in the Coq6 RATIONAL homology model.*

Even when a residue is known to form part of a tunnel bottleneck, its orientation can change over the course of the molecular dynamics simulation because of side-chain rotation. **Figure 4.14** shows us the case of L382 and P249 in the RATIONAL Coq6 model. Using the distance measurement function in VMD, we must select specific atoms of each bottleneck residue in order to measure tunnel diameter. However, L382 is a branched sidechain with rotation possible about the CB-CG bond, which is topologically upstream of the terminal carbons, CD1 and CD2. Physically, it makes more sense to select these terminal carbon atoms for distance measurements, since they are closest to the tip of the sidechain. We do not consider the terminal CD1 and CD2 bound hydrogens for this selection because their positions will not be reproduced by AutoDock Vina's molecular representation (which represents only polar hydrogens). However, because rotation is possible about the CB-CG bond, selecting only one sidechain atom, (CD1 or CD2), is likely to give an artefactual measurement of the tunnel diameter. For example, if we measure the distance between L382_CD1 P249_CG, we may perform this calculation on a frame where the L382 sidechain has rotated, interchanging the positions of CD1 and CD2. If we plot the L382_CD1-P249_CG distance over the course of the trajectory, we will see a maximum in the distance curve, which might be

tempting to interpret as a frame where the tunnel is open. However, if rotation interchanges the positions of L382 CD1 and CD2 then the tunnel will actually be blocked in a similar fashion. Therefore we would like to select an atom that will better represent the spatial position of the sidechain tip. This leads us to backtrack up the molecular topology of the sidechain to the CG carbon of the L382 sidechain. This is a good selection because of it is immediately upstream of the terminal CD1 and CD2 atoms, yet it is not subject to rotational degeneracy itself. However, while selecting such an atom eliminates rotational degeneracy because of the branched side-chain, it means we must add diameter corrections to represent the physical volume of the CD1 and CD2 atoms. This is developed in the following section.

4.4 van der Waals radius corrections

The purpose of measuring tunnel diameters over the course of molecular dynamics is to select frames from molecular dynamics likely to be compatible with substrate docking. Therefore, the van der Waals radius correction we will develop is designed to represent atomic dimensions as they are represented in AutoDock Vina. However in AutoDock Vina, the sizes of atoms are specified as spheres of fixed radius, whereas in the molecular dynamics force field, they are defined by a potential energy function. This means that the inter-atomic distances recorded from molecular dynamics simulations do not have a fixed value; rather, they vary harmonically over an interatomic distance range specified by force field parameters. This means that sometimes the interatomic distances will be smaller than expected based on fixed van der Waals radii. The practical result is that when we apply fixed distance van der Waals radius corrections to the inter-atomic distances recorded from molecular dynamics trajectories, we can obtain tunnel diameter values that are negative for some frames.

An important point to make in this section is the differing representations of the same molecular structures by the different programs used for each modeling task. MODELLER generates all-atom protein models including non-polar hydrogens, as does the force-field chosen for our molecular dynamics simulations (AMBER99-SB-ILDN) also uses an all atom representation. However, the method we will use for molecular docking, AutoDock Vina, only represents polar hydrogens explicitly. As part of this simplification, carbons bearing non-polar hydrogens are represented as particles with larger van der Waals radii than those described in the molecular dynamics force field. The practical consequence of this is that we must analyze our all-atom trajectories from the point of view of AutoDock Vina's united-atom representation. Concretely, this means that we must apply van der Waals radius corrections based on Vina's specific united atom van der Waals radius. These are given in **Table 4.1** below.

TABLE 4.1: *VDW radii for selected atoms as represented by the Amber forcefield in molecular dynamics and by AutoDock Vina in molecular docking.*

Atom	Amber VDW radius (Å)	Vina VDW radius (Å)
C	1.7	2
N	1.625	1.75
O	1.5	1.6
H	1	1

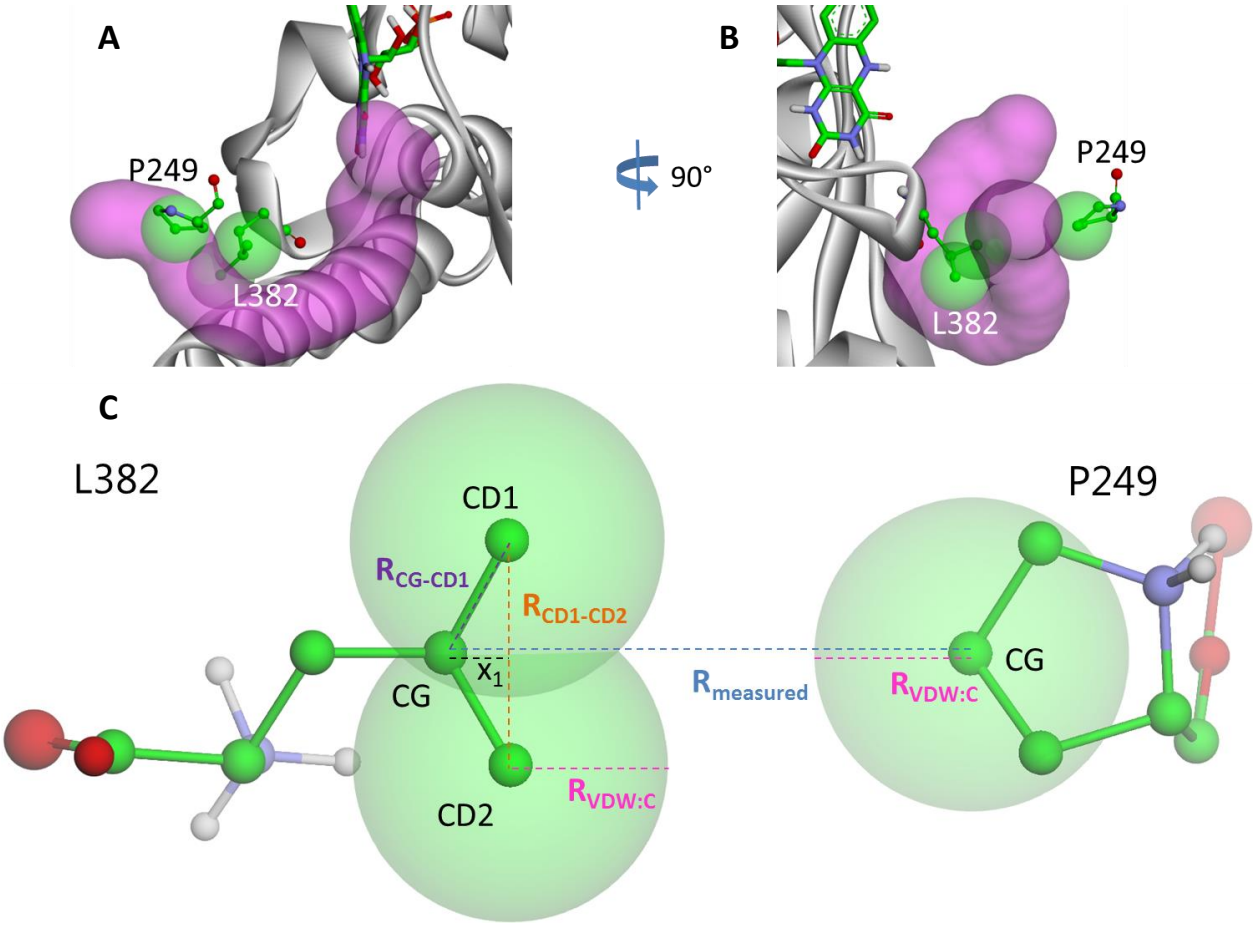

FIG 4.15: *The re-face tunnel 1 of the RATIONAL Coq6 model (purple volume) A) shown in the same orientation as Figure 12, B) rotated by 90 degrees, C) an illustration of the distances used to calculate the effective diameter of the tunnels.*

We will use *re* face tunnel 1 of the RATIONAL Coq6 model to develop an example of the van der Waals radius correction for calculating the effective diameter of the tunnel. For this tunnel, we have identified P249 and L382 as bottleneck residues. However, as explained in **Figure 4.14**, we selected specific atoms to measure the tunnel diameter over the course molecular dynamics: the CG atom of L382 and the CG atom of P249. However, we must subtract the van der Waals radii of the intervening atoms in order to get a better estimate of the tunnel's actual diameter. Since we are assessing the diameter in order to select enzyme conformations for substrate docking with AutoDock Vina, we will use Vina's van der Waals radii to make the corrections. The correction term is composed of the Vina van der Waals radii for the potentially contacting atoms (L382 CD atoms and the P249 CG atom), plus the x_1 distance, which is the forward displacement of the L382 CD atoms (the atoms which could potentially contact P249) relative to the L382 CG atom. In this case, the x_1 distance is calculated by the Pythagorean theorem to be 0.6 Å.

Another important point is that neither the tunnel nor the substrate have circular cross sections. This means that both geometric constructs, the tunnel and the substrate, have cross-sectional geometries

that are better described by “diameters” on multiple axes, therefore **describing the minimum passable diameter of the tunnel with a single number is not entirely accurate, and may miss traversable tunnel conformations.** Moreover, the tunnel and the substrate can have different relative orientations, which means that making a comparison between the effective diameters of the substrate and the tunnel requires describing their relative orientations. This introduces a lot of unnecessary complications in geometry, since a more detailed description of relative tunnel-substrate orientations would require many more terms: at least one dihedral angle for every degree of rotational freedom in the substrate and in the enzyme bottleneck sidechains. Even if we did develop such a detailed representation, it would be of limited use because we do not know *a priori* (before docking) what relative orientations correspond to traversable conformations of the tunnel.

We can avoid all of this with the equivalent of a limited coarse-graining, which has the goal of simply describing the maximum separation between the bottleneck residue sidechains. **Since we have no detailed *a priori* knowledge of enzyme conformations compatible with passage of the substrate**, we must set the most general (yet accurate) criterion possible: **we will look for the maximum effective diameter at each tunnel bottleneck**. This measurement contains the fewest underlying hypotheses and geometrical constructions. Even though the tunnel and the substrate have several diameters along several axes, selecting frames with maximum bottleneck diameters gives us the best chance to allow passage of the substrate, even in its least favorable orientation.

In the following section, we will develop and apply these corrections to each of the tunnels in each of the models.

4.5 *re* face tunnel 1

I-TASSER Coq6 model

Figure 4.16A shows the *re* face tunnel 1 alone, from the same view as in **Figure 4.10**. **Figure 4.16B** shows the same tunnel, rotated 90 degrees so that we are looking through the *re* face of the enzyme, more clearly presenting the bottleneck formed by the sidechain of the residues I250 and L382. **Figure 4.16C** illustrates these sidechains in a head-to-head orientation, identifying the specific atoms used to monitor the diameter as well as their VDW radii. In this bottleneck the sidechains of both residues point inwards to the tunnel lumen, and can visit conformations where they are diametrically opposite each other. Therefore we selected the CG atom from the L382 sidechain and the CD atom from the tip of the I250 sidechain as points between which to measure the tunnel bottleneck diameter, as shown in **Figure 4.16C**.

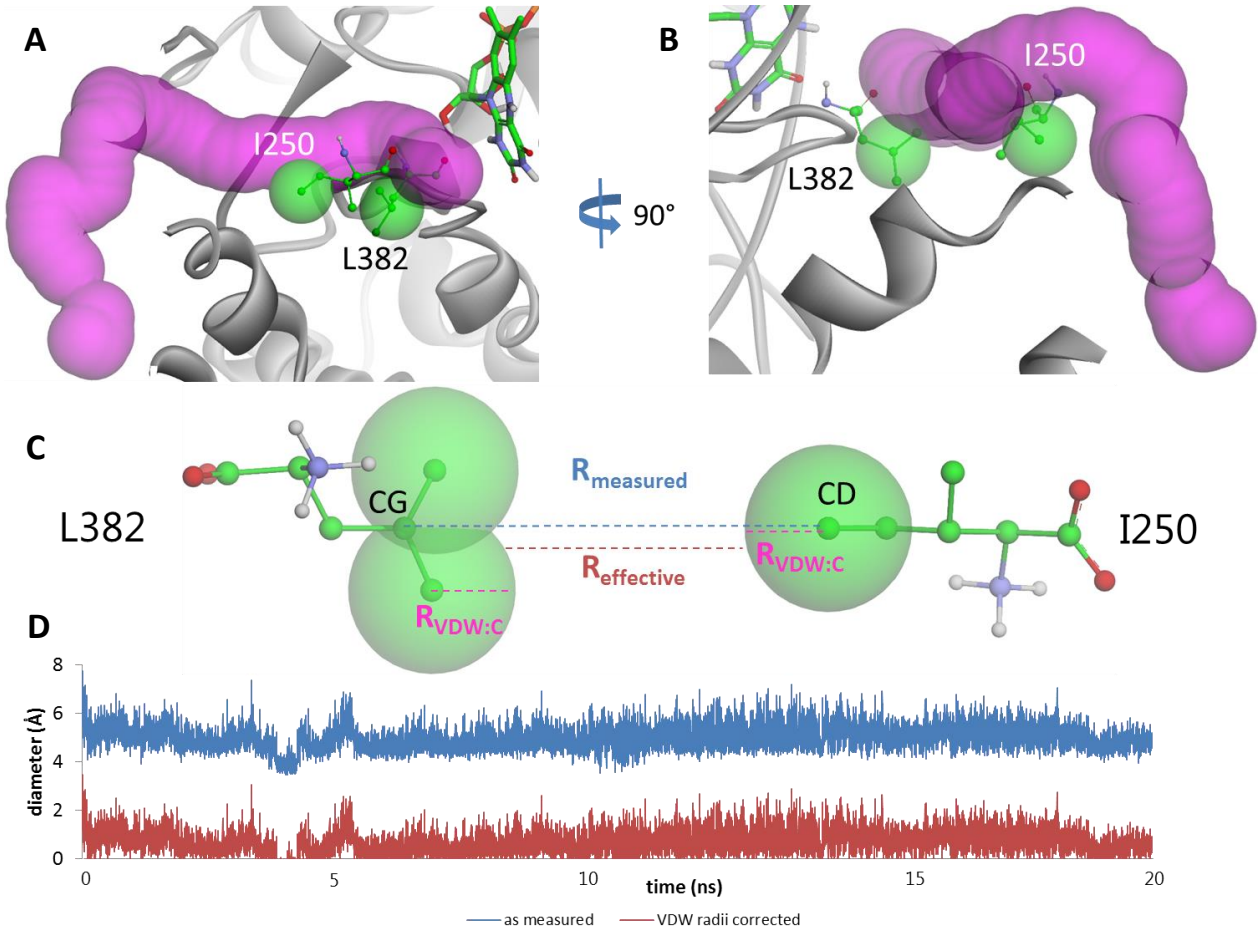

FIG 4.16: *A) ITASSER Coq6 model re face tunnel 1 bottleneck residue identification, seen from same view as in Figure 4.4, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. D) A plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

We then subtract the VDW radii of these atoms (as listed in **Table 4.1**) from the measured distance ($R_{measured}$) to find the effective diameter ($R_{effective}$) of the tunnel at this point. **Figure 4.16D** shows the evolution of these distances (as measured and after VDW radius correction) over the course of the simulation. The value descends to an average of 1.1Å during the latter half of the simulation, a diameter too small to allow passage of the substrate model in any orientation. However, the trace does show some maxima that can reach 3Å. In order to explicitly test the ability of this tunnel to allow passage of the substrate, we will select conformations corresponding to these maxima for subsequent docking studies.

ROBETTA Coq6 model

Figure 4.17A shows the *re* face tunnel 1 alone, from the same view as in **Figure 4.11**. **Figure 4.17B** shows the same tunnel, rotated 90 degrees so that we are looking through the *re* face of the enzyme, more clearly presenting the bottleneck formed by the sidechains of residues S265 and L382. **Figure 4.17C** shows an illustration of these sidechains in a head to head orientation, identifying the specific atoms

used to monitor the diameter as well as their VDW radii. In this bottleneck the sidechains of both residues point inwards to the tunnel lumen and can be diametrically opposed to each other.

The rotational degeneracy possible for the L382 sidechain through rotation about the CB-CG bond (as shown in **Figure 4.14**) means that selection of either of the sidechain's terminal CD1 or CD2 atoms will give a misleading index of tunnel diameter. Therefore we select the next atom upstream in the topology, the CG atom. The S265 sidechain, which is unbranched, presents a simpler geometrical case with no rotational degeneracy possible, allowing us to select the terminal oxygen atom as a distance measurement point.

We then subtract the VDW radii of these intervening atoms (as listed in **Table 4.1**) from the measured distance ($R_{measured}$) to find the effective diameter ($R_{effective}$) of the tunnel at this point. The selection of the L382 CG atom means that we must add a correction term for the forward distance between the CG atom and the plane of the CD1-CD2 atom pair, as indicated in **Figure 4.17**. This distance is easily calculated with the Pythagorean theorem, and then simply added to the Vina VDW radius of a carbon atom. **Figure 4.17D** shows the evolution of the diameter over the course of the molecular dynamics simulation. The tunnel diameter stabilizes to an average value of 1.48Å over the latter half of the simulation. This is also too small to allow passage of the substrate. However, in order to allow for the possibility that this tunnel may be transiently passable (as some conformations display tunnel diameters approaching 4Å) we will still select maxima from this curve for sampling enzyme conformations for substrate docking.

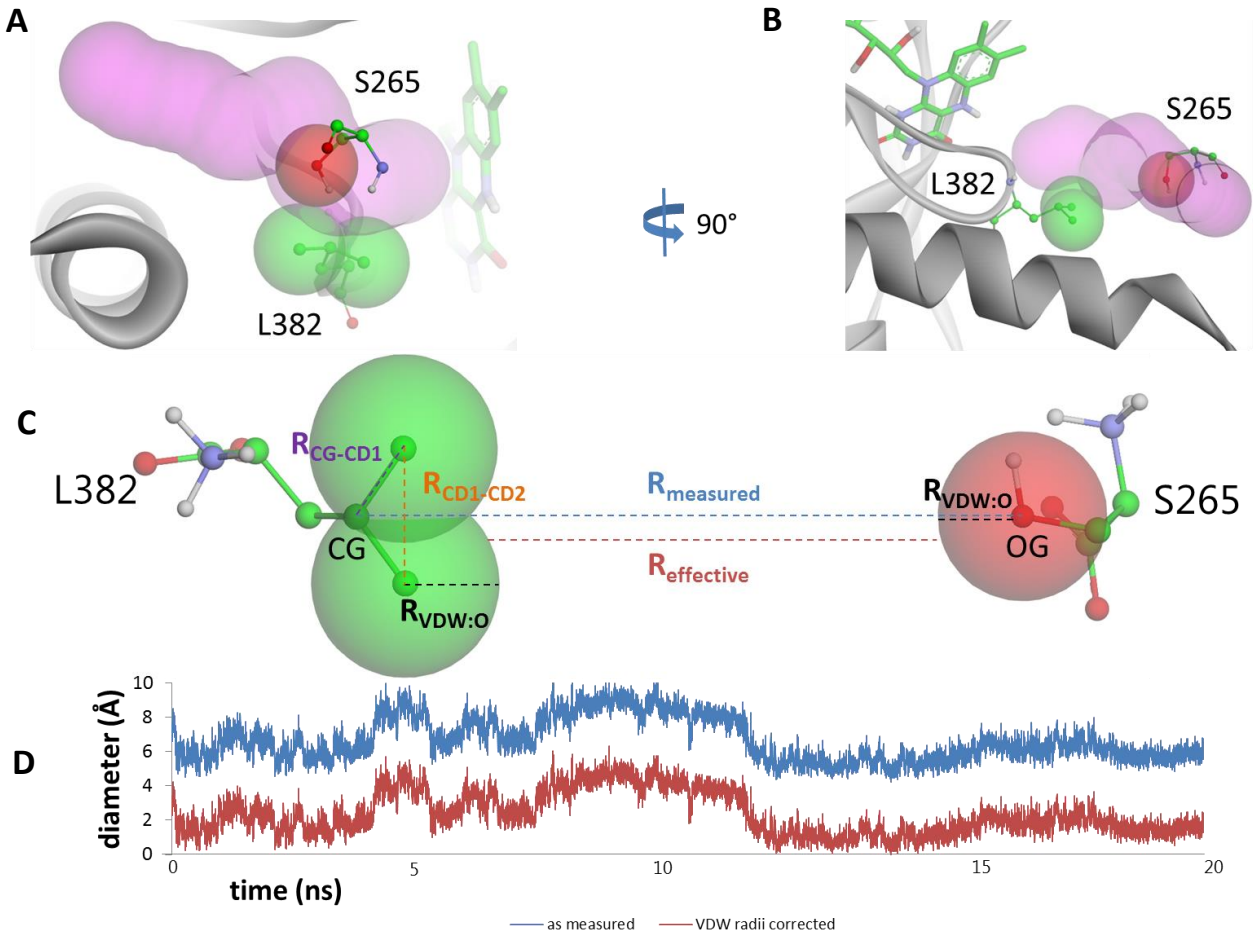

FIG 4.17: *A) ROBETTA Coq6 model re face tunnel 1 bottleneck residue identification, seen from same view as in Figure 4.11, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. D) A plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

RATIONAL Coq6 model

Figure 4.18A shows the *re* face tunnel 1 alone, from the same view as in **Figure 4.12**. **Figure 4.18B** shows the same tunnel, rotated 90 degrees so that we are looking through the *re* face of the enzyme, more clearly presenting the bottleneck formed by the sidechains of residues P249 and L382. **Figure 4.18C** shows an illustration of these sidechains in a head to head orientation, identifying the specific atoms used to monitor the diameter as well as their VDW radii. In this bottleneck the sidechains of both residues point inwards to the tunnel lumen and can be diametrically opposite each other.

Once again the L382 sidechain presents a rotational degeneracy in the sidechain tip, so we will select its CG atom as an anchor point for measurement. From the P249 residue we select the CG atom, since it is at the apex of the cyclized sidechain. The correction factor for the L382 sidechain is calculated as previously described. The effective diameter of the tunnel stabilizes around a value of 3.4Å, larger than

the other tunnels examined so far. As for all tunnels in all models, we will still select enzyme conformations corresponding to maxima in the distance curve for subsequent substrate docking calculations.

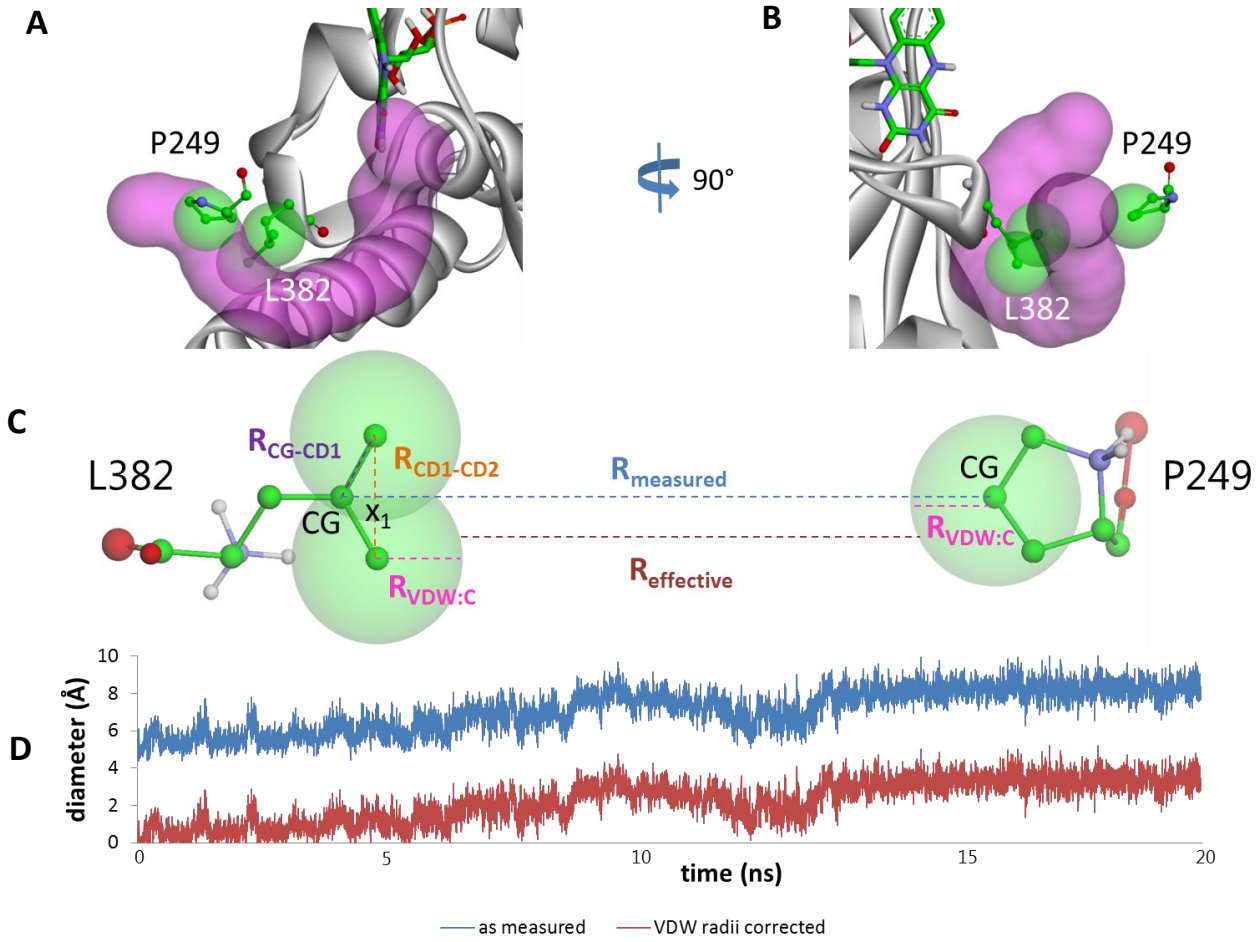

FIG 4.18: *A) RATIONAL Coq6 model re face tunnel 1 bottleneck residue identification, seen from same view as in Figure 4.12, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. The forward distance between the L382 gamma carbon (CG, where the distance is measured) and the atoms of the sidechain tip (delta carbons CD1 and CD2) is calculated with the Pythagorean theorem. D) A plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

4.6 *re* face tunnel 2

I-TASSER Coq6 model

Figure 4.19A shows the *re* face tunnel 2 alone, from the same view as in **Figure 4.10**. **Figure 4.19B** shows the same tunnel, rotated 90 degrees so that we are looking through the *re* face of the enzyme, more clearly presenting the bottleneck formed by the sidechains of residues M434 and L452. **Figure 4.19C** shows an illustration of these sidechains in a head to head orientation, identifying the specific atoms used to monitor the diameter as well as their VDW radii. In this bottleneck the sidechains of both residues point inwards to the tunnel lumen and can be diametrically opposite each other. Once again we have a leucine residue at this bottleneck (L452), so we will select its CG atom as the measurement point. The M434 sidechain is linear, so will select its terminal CE as the other measurement point. The VDW radius corrections for L452 are developed as in prior cases; for M434 it is simply the Vina VDW radius for a carbon atom. The effective diameter of the tunnel stabilizes around a value of 1.9Å.

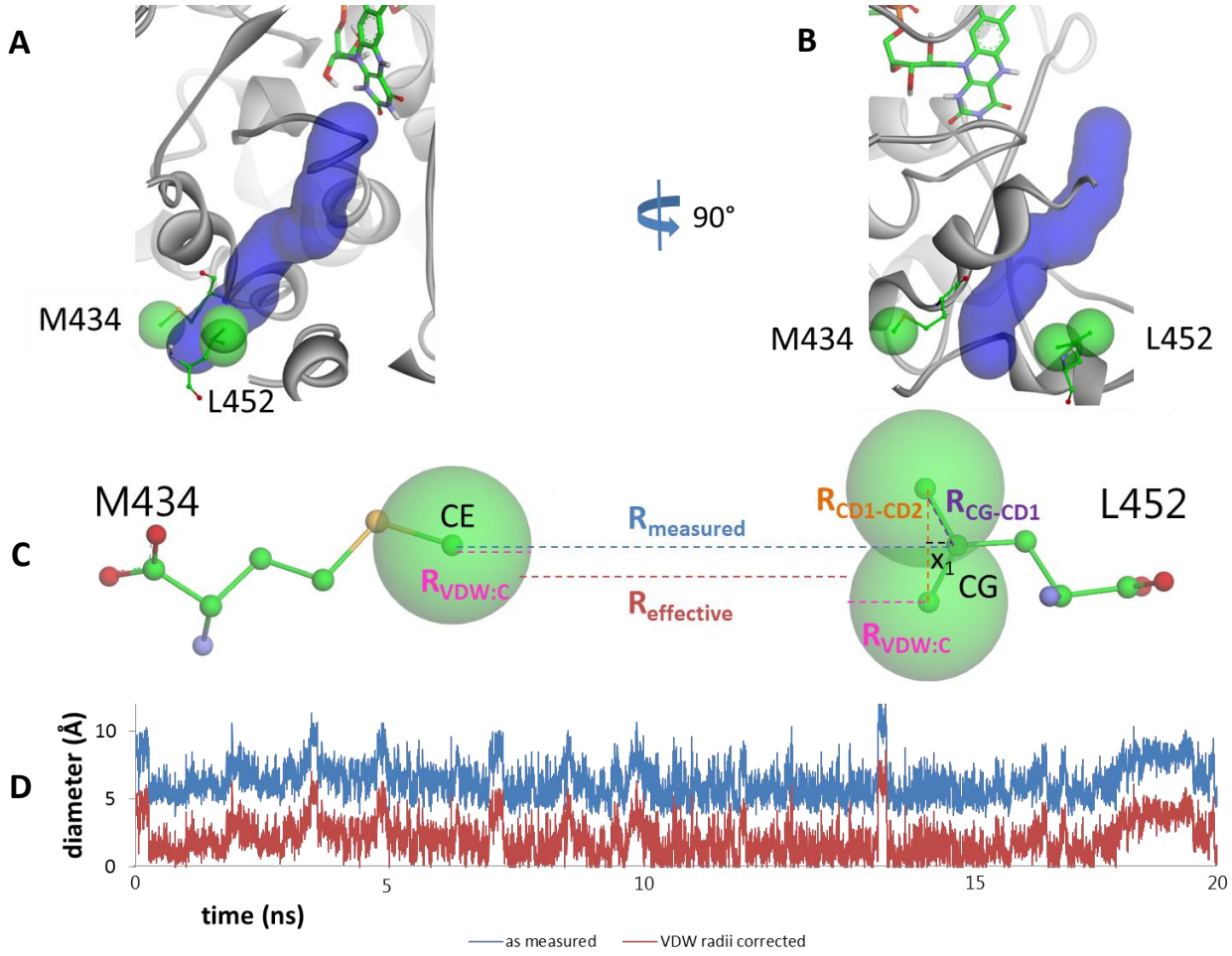

FIG 4.19: *A) ITASSER Coq6 model re face tunnel 2 bottleneck residue identification, seen from same view as in Figure 4.10, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. The forward distance between the L452 gamma carbon (CG, where the distance is measured) and the atoms of the sidechain tip (delta carbons CD1 and CD2) is calculated with the Pythagorean theorem. D) Plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

ROBETTA Coq6 model

Figure 4.20A shows the *re* face tunnel 2 alone, from the same view as in **Figure 4.11**. **Figure 4.20B** shows the same tunnel, rotated 90 degrees so that we are looking through the *re* face of the enzyme, more clearly presenting the bottleneck formed by the sidechains of residues Q242 and L474. **Figure 4.20C** shows an illustration of these sidechains in a head to head orientation, identifying the specific atoms used to monitor the diameter as well as their VDW radii. In this case the bottleneck is formed by the sidechain oxygen and nitrogen atoms of Q242 and the terminal CD carbons of the L474 sidechain.

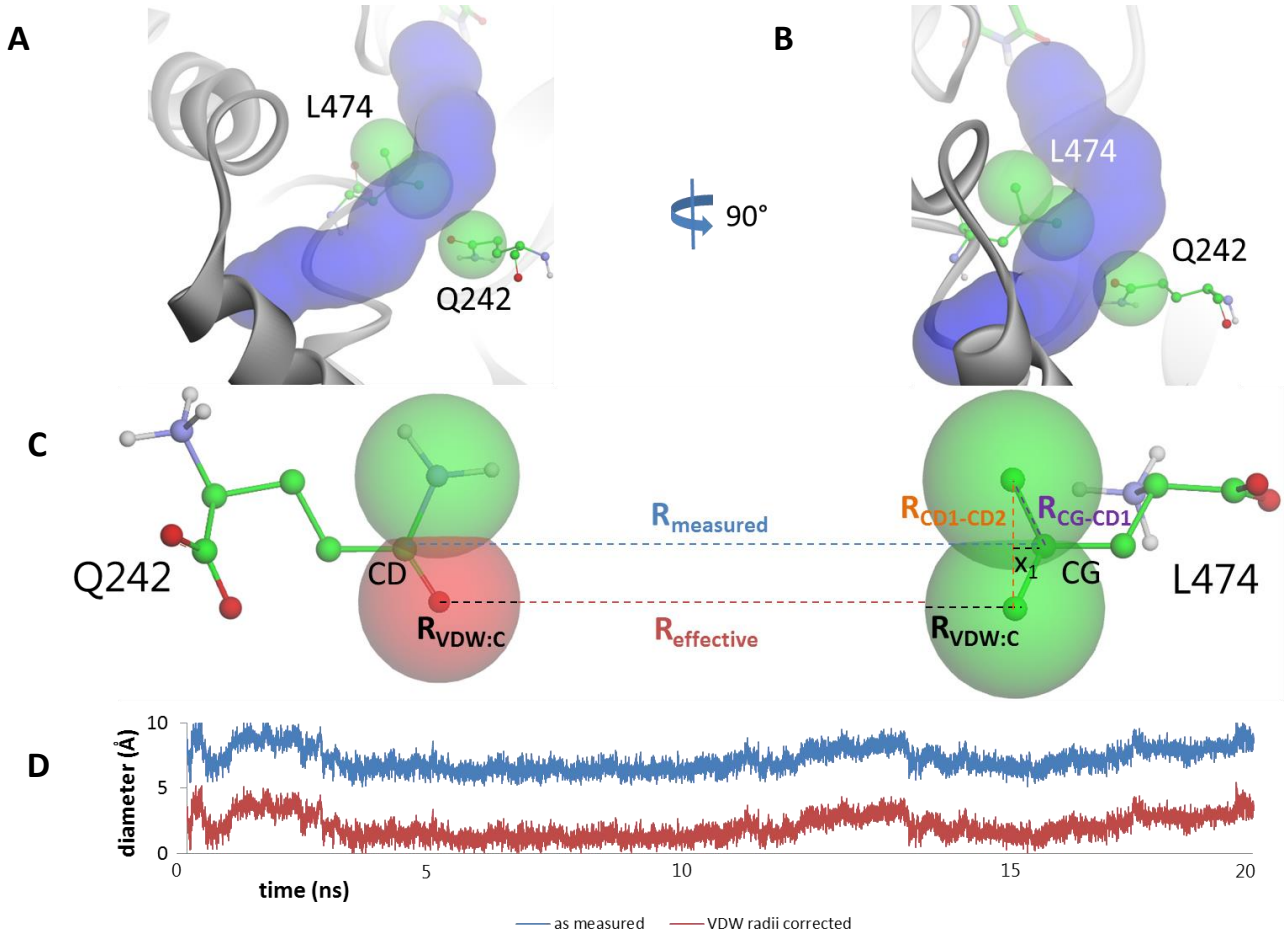

FIG 4.20: *A) ROBETTA Coq6 model re face tunnel 2 bottleneck residue identification, seen from same view as in Figure 4.11, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. D) A plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

RATIONAL Coq6 model

Figure 4.21A shows the *re* face tunnel 2 alone, from the same view as in **Figure 4.12**. **Figure 4.21B** shows the same tunnel, rotated 90 degrees so that we are looking through the *re* face of the enzyme, more clearly presenting the bottleneck formed by the sidechain of the residues L441 and L474. **Figure .21C** shows an illustration of these sidechains in a head to head orientation, identifying the specific atoms used to monitor the diameter as well as their VDW radii and sidechain geometry correction terms as developed for the leucine sidechain.

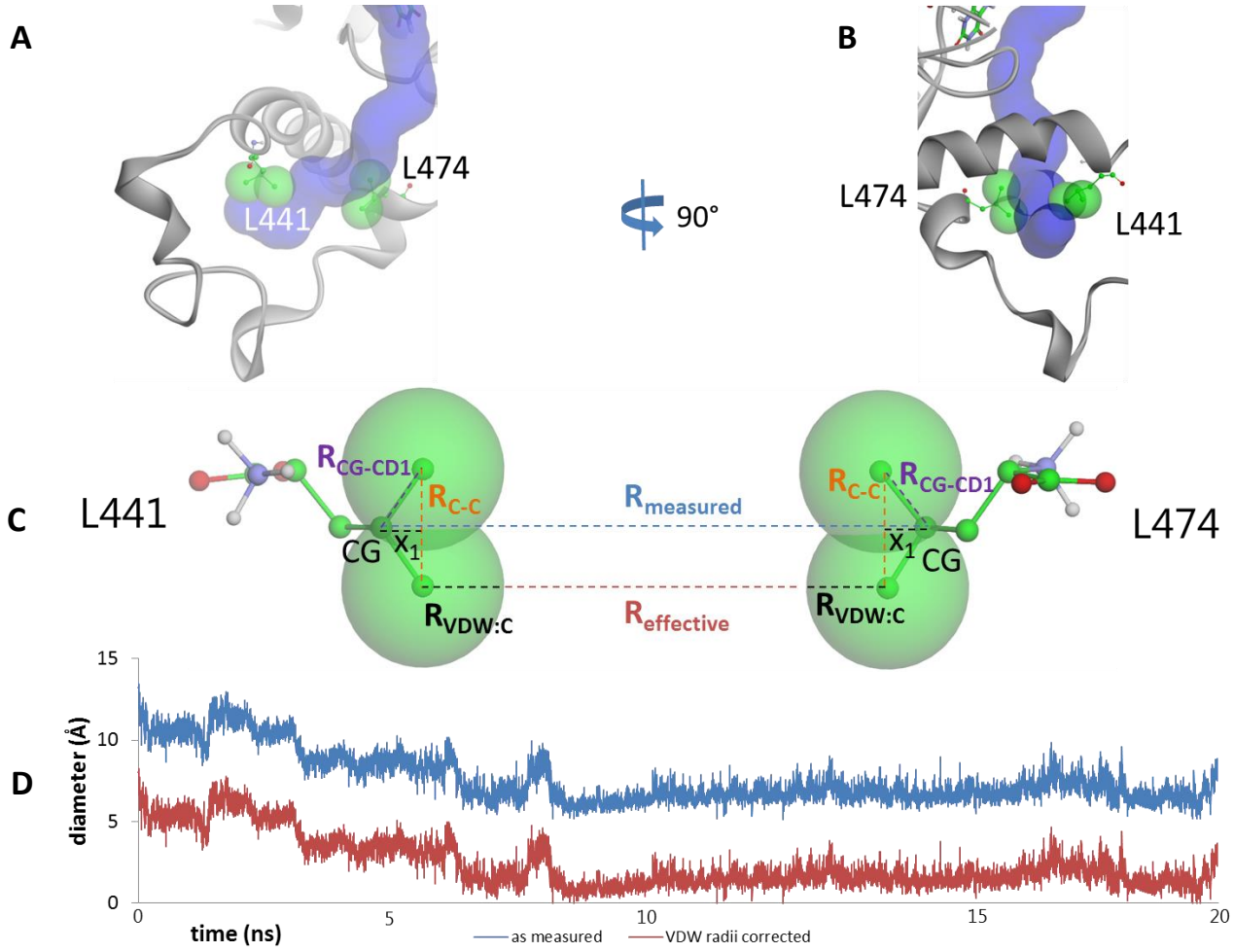

FIG 4.21: *A) RATIONAL Coq6 model re face tunnel 2 bottleneck residue identification, seen from same view as in Figure 4.12, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. The forward distance between the leucine gamma carbon (CG, where the distance is measured) and the atoms of the sidechain tip (delta carbons CD1 and CD2) is calculated with the Pythagorean theorem. D) Plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

The *re* face 2 tunnels also show distinctly different behavior among the models. In the I-TASSER model the tunnel diameter converges to an average value of 1.94 Å, suggestive of a tunnel that is essentially closed. In the ROBETTA model, this tunnel starts off as being very large (13.7 Å), but collapses to an average value of 6.02 Å during the latter half of the simulation, which could be large enough to permit passage of the substrate, which has a diameter of 4 Å in one dimension. Similar behavior is observed for this tunnel type in the RATIONAL model, which starts at a diameter of 8 Å but collapses to an average diameter of 1.53 Å by simulation end.

4.7 *si* face tunnel 1

I-TASSER Coq6 model

Figure 4.22A shows the *si* face of the tunnel alone, from the same view as in **Figure 4.10**. **Figure 4.22B** shows the *si* face tunnel alone, from the same view as in **Figure 4.10**. **Figure 4.22C** shows the same tunnel, rotated 90 degrees so that we are looking at the *si* face of the enzyme, more clearly presenting the bottleneck formed by the sidechain of residue P256 and the backbone of I136. **Figure 4.22C** shows a representation of the relative orientation of these sidechains. Therefore we will use the VDW radii for these atoms in making the corrections to calculate the effective diameter. The average diameter of this tunnel is around 3 Å.

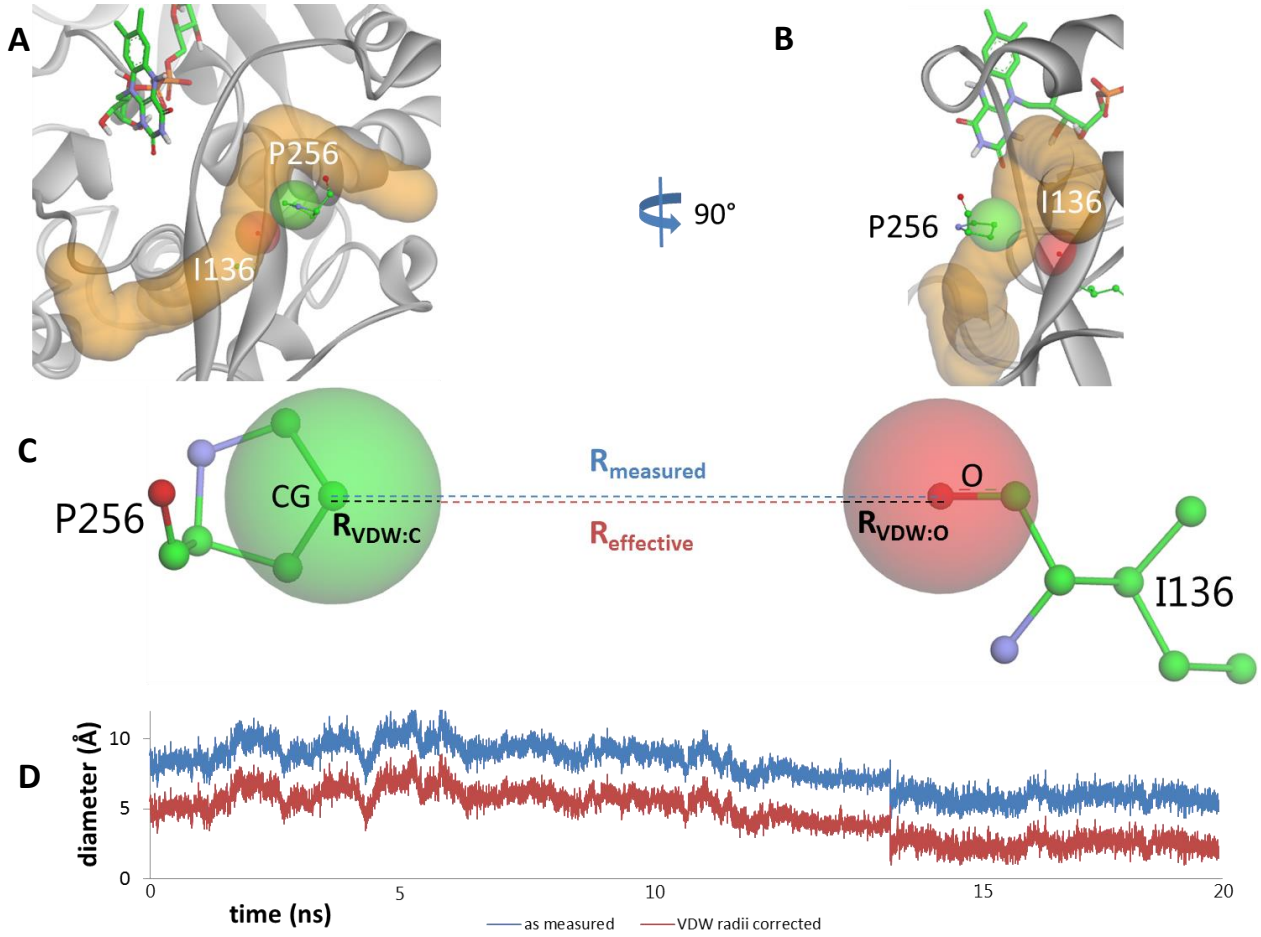

FIG 4.22: *A) I-TASSER Coq6 model si face tunnel bottleneck residue identification, seen from same view as in Figure 4.10, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. D) A plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

ROBETTA Coq6 model

Figure 4.23A shows the *si* face tunnel alone, from the same view as in **Figure 4.11**. **Figure 4.23B** shows the same tunnel, rotated 90 degrees so that we are looking at the *si* face of the enzyme, more clearly presenting the bottleneck formed by the sidechain of residues V85 and S228. **Figure 4.23C** shows an illustration of these sidechains in a head to head orientation, identifying the specific atoms used to monitor the diameter as well as their VDW radii and sidechain geometry correction terms as developed for the leucine sidechain. In this case both sidechains are branched, therefore we will develop the VDW correction radii in the same manner as introduced for the leucine residue. The average diameter of the tunnel stabilizes around 6.02 Å, which might be large enough to allow passage of the substrate, on its smallest dimension.

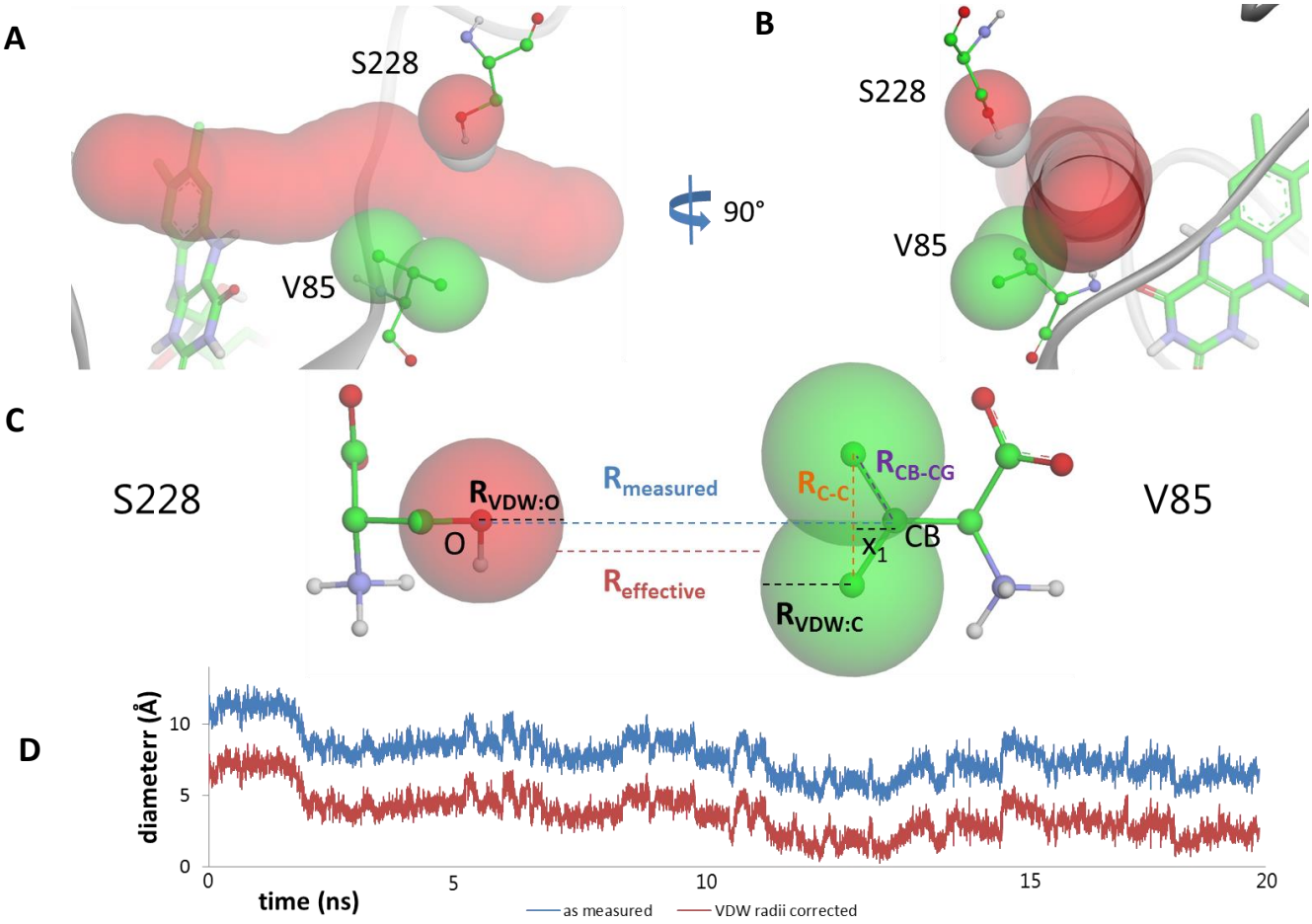

FIG 4.23: *A) ROBETTA Coq6 model si face tunnel bottleneck residue identification, seen from same view as in Figure 4.11, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. The forward distance between the valine beta carbon (CB, where the distance is measured) and the atoms of the sidechain tip (gamma carbons CG1 and CG2) is calculated with the Pythagorean theorem. D) Plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

RATIONAL Coq6 model

Figure 4.24A shows the *si* face tunnel alone, from the same view as in **Figure 4.12**. **Figure 4.24B** shows the same tunnel, rotated 90 degrees so that we are looking at the *si* face of the enzyme, more clearly presenting the bottleneck formed by the sidechain of residues E137 and M255. **Figure 4.24C** shows an illustration of these sidechains in a head to head orientation. E137 is a branched sidechain, and we develop this VDW correction as for leucine. The relative orientation of the M255 can place the terminal CE carbon in contact with E137, therefore we select the CE atom for adding the VDW correction on this side of the bottleneck. The average diameter of this tunnel stabilizes around 1.9 angstroms, although it is capable of opening to 6.7 Å. In order to test the possibility of substrate passage, we will choose the strategy of selecting enzyme conformations corresponding to maxima in the diameter plot.

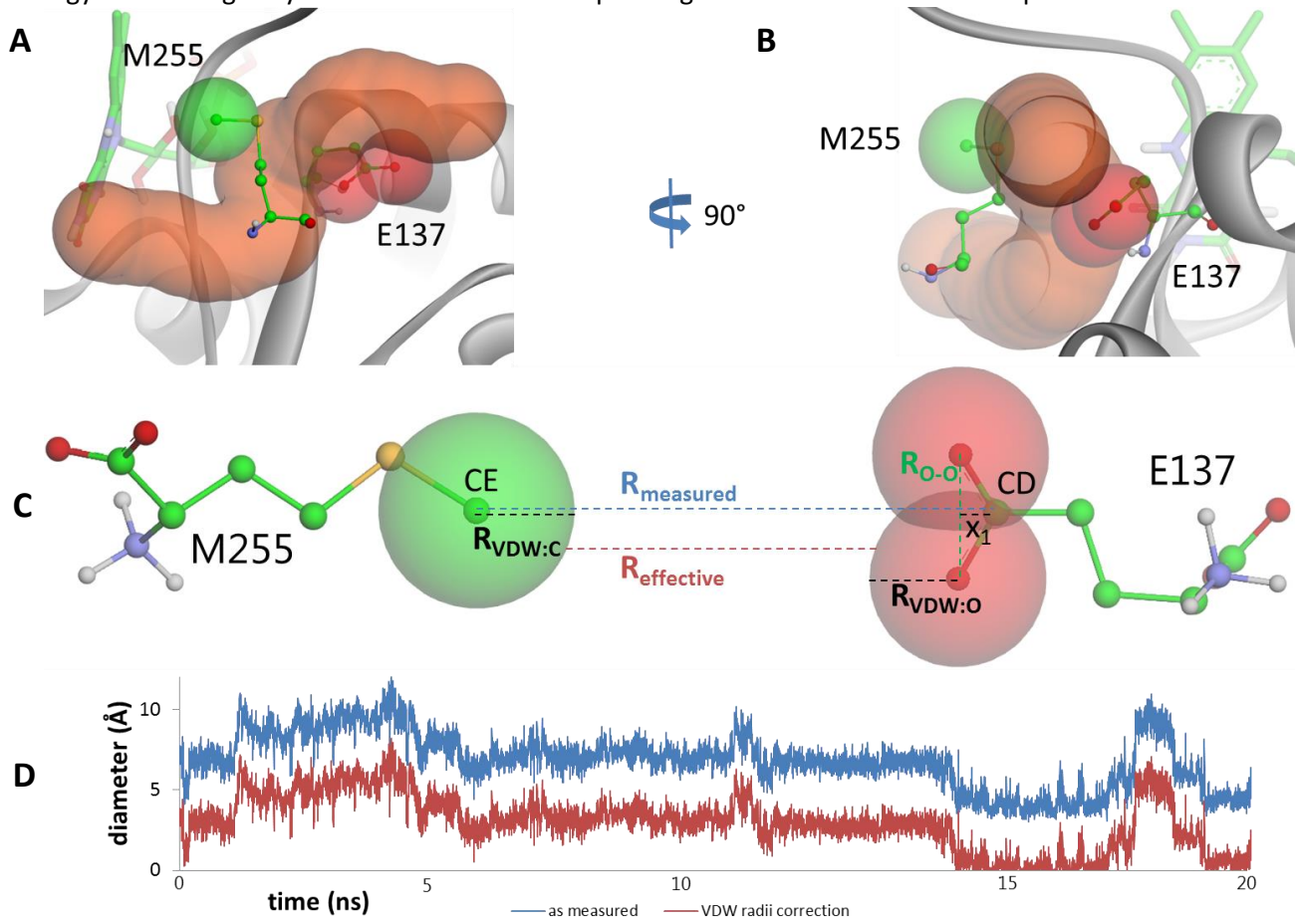

FIG 4.24: *A) RATIONAL Coq6 model si face tunnel bottleneck residue identification, seen from same view as in Figure 4.12, B) rotated 90 degrees and seen from the re face. C) Diagram illustrating the inter-nuclear distance $R_{measured}$ tracked over MD for this tunnel bottleneck diameter and the van der Waals radius corrected $R_{effective}$ approximating the diameter available to a substrate in subsequent docking attempts. The forward distance between the glutamate delta carbon (CD, where the distance is measured) and the atoms of the sidechain tip (epsilon carbons CE1 and CE2) is calculated with the Pythagorean theorem. D) Plot of the $R_{measured}$ and $R_{effective}$ over the course of MD.*

4.8 Conclusions: comparison of the 3 tunnel types

The analysis of the Coq6 models reveals the existence of three types of tunnels leading from the protein surface to the active site. Two of these tunnels pass through the *re* face of the enzyme and one of them passes through the *si* face, converging onto a volume directly in front of the FAD's isoalloxazine ring. One of these tunnels, *re* face tunnel 1, is composed of residues that evolutionarily conserved in the Coq6 family of proteins. Residue conservation around a geometric feature proximal to the active site is suggestive of a possible substrate access region. Therefore, we characterized the tunnels by their diameter over molecular dynamics trajectories based on the hypothesis that these might be used to permit passage of the substrate to the active site. Most of the tunnels collapsed to average diameters of less than 4 Å, with the exception of the *re* face 1 tunnel in the RATIONAL model.

The *re* face 1 tunnels behave differently among the three homology models. In the I-TASSER model the tunnel diameter at the I250-F439 bottleneck converges to an average value of 1.09 Å in the second half of the simulation. This is very small and likely to be too small to permit passage of the substrate (whose largest diameter at the aromatic head is 8.37 Å), even if we consider maximum values of the tunnel diameter, which only reach 4.5 Å. In the ROBETTA model this tunnel displays similar behavior, stabilizing around an average diameter of 1.48 Å. This tunnel shows a very different behavior in the manually curated multi-template model. It begins with a much higher initial diameter of about 5 Å, converges to an average value of 8.2 Å in the latter half of the simulation, and shows much smaller diameter fluctuations over this range.

The *re* face 2 tunnels also show distinctly different behavior among the models. In the I-TASSER model the tunnel diameter converges to an average value of 1.9 Å, suggestive of a tunnel that is essentially closed. In the ROBETTA model, this tunnel starts off as being very large (13.7 Å), but collapses to an average value of 6.02 Å during the latter half of the simulation. A similar behavior is observed for this tunnel type in the RATIONAL model, which starts at a diameter of 8 Å, but collapses to an average diameter of 1.53 Å by simulation end.

Finally, the *si* face tunnels show more variability. In the I-TASSER model this channel starts off small with a diameter of 0.28 Å and finished with an average diameter of 1.38 Å. In the ROBETTA model this tunnel starts off with a diameter of 6.09 Å and finishes with an average diameter of 2.58 Å. A similar collapse of this tunnel is observed for the RATIONAL model.

We will now test the ability of these tunnels to permit the passage of the substrate to the active site by molecular docking of Q biosynthesis intermediates into enzyme conformations selected from molecular dynamics trajectories on the basis of maximum diameter of potential substrate access tunnels.

5. Substrate access channel characterization

So far we have tested the ability of the AutoDock Vina docking program to find the active site without any extra information, and we have tested the effect of polyprenyl chain length on the placement of the active site. However, we have presented these preliminary tests only on the RATIONAL Coq6 model. We will now proceed to a more systematic characterization and testing of the tunnels identified in the previous chapter.

As described in this chapter's introduction, we will use the technique of ensemble docking, wherein we dock our model substrate into several conformations of the same receptor. The molecular dynamics simulations we ran for structural stability testing have also computed many different conformations of

the enzyme for us: 10 000 frames covering 20 nanoseconds. This is too many conformations for us to screen with docking in a reasonable amount of time, and even if it were feasible, the results of 10 000 docking jobs would be very time consuming to analyze. Since we know that Coq6 binds the substrate from *in vivo* experimental knowledge, our remaining task is to select Coq6 conformations that are most compatible with substrate binding if we wish to observe a physically plausible result from substrate docking calculations.

Thus far we have characterized each of the three tunnels of Coq6 by its diameter at its narrowest point. We reason that if the tunnel is not traversable at its maximum diameter, it is not traversable at all, and therefore not a functional substrate access channel. For recall, we have the *re* face tunnel 1 (always colored in purple in our figures), the *re* face tunnel 2 (always colored in blue), and the *si* face tunnel (always colored in red). Calculating a diameter for every frame of the molecular dynamics simulations allows us to select specific conformations from each model's trajectories where each tunnel has a maximum diameter. Therefore our first criteria for selecting frames from the three Coq6 model trajectories will be the maximal diameter for each tunnel. Despite the fact that most of the tunnels collapsed to low average diameters by the end of the simulations, they still display thermal fluctuations and may be able to visit conformations where they are wide enough to permit passage of the substrate to the active site. We will resume the results of the tunnel diameter characterization with the following **Figure 4.25**.

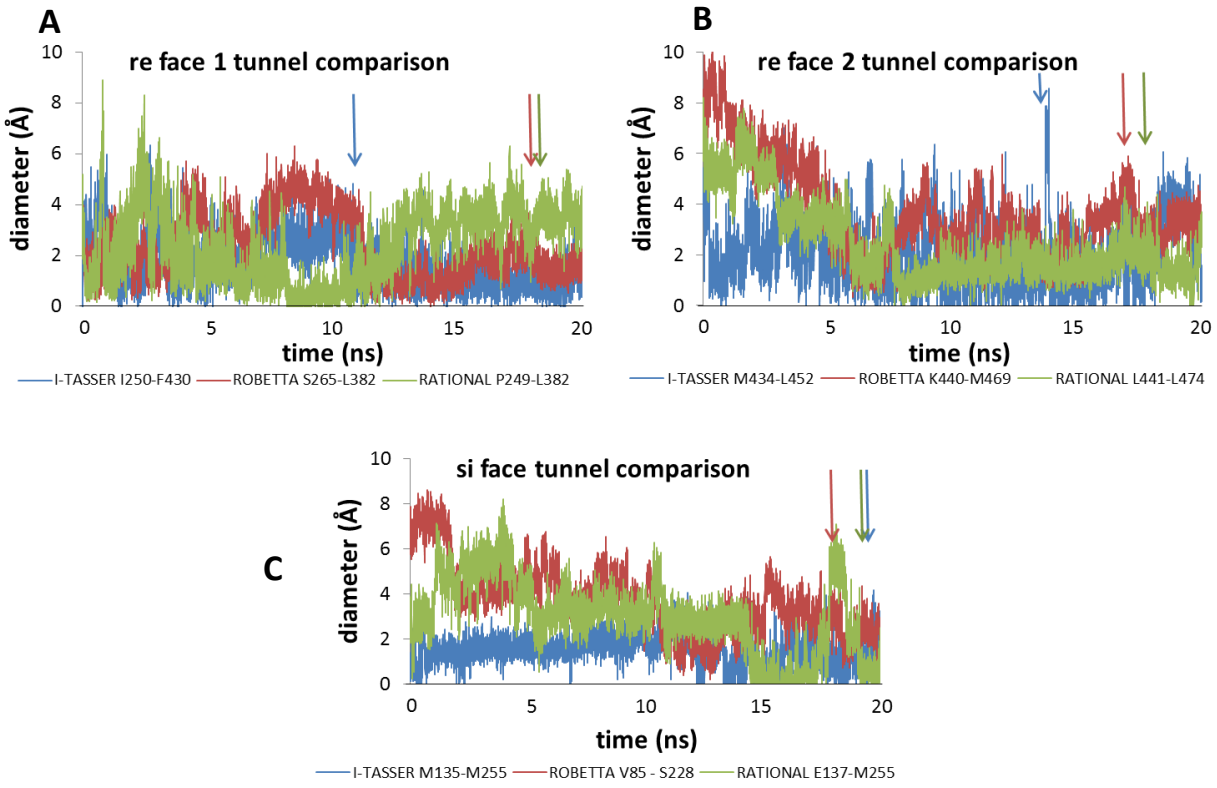

FIG 4.25: *Tunnel bottleneck diameter comparison by tunnel type as a function of time over the 20ns MD simulations. Tunnel diameters from the I-TASSER model are shown in blue; from the ROBETTA model, in red; from the RATIONAL model, green. Frames sampled for substrate docking are indicated by color coded arrows.*

Calculating a diameter for every frame allows us to sort the frames of the trajectory by their diameter, enabling us to select the frames with the largest tunnel diameter for docking attempts. For each tunnel in each model, we will select the best conformations for docking, presenting the docking results below.

5.1 Round 1 of docking: Channel traversability screening

The results from docking our model substrate 4-HP6 (3-hexaprenyl-4-hydroxyphenol) for the three models are presented below. For these runs we use the same type of docking box shown in **Figure 4.23**, with minor adjustments to account for the changing shapes and locations of the tunnels. While all three tunnel types are therefore included in the docking box for these runs, the tunnel specificity for each docking run is given by the selection of specific frames from each docking run on the criterion of tunnel diameter.

5.1.1 Substrate docking into the I-TASSER Coq6 model

The results from docking our model substrate (3-hexaprenyl-4-hydroxyphenol) into the I-TASSER model are presented below.

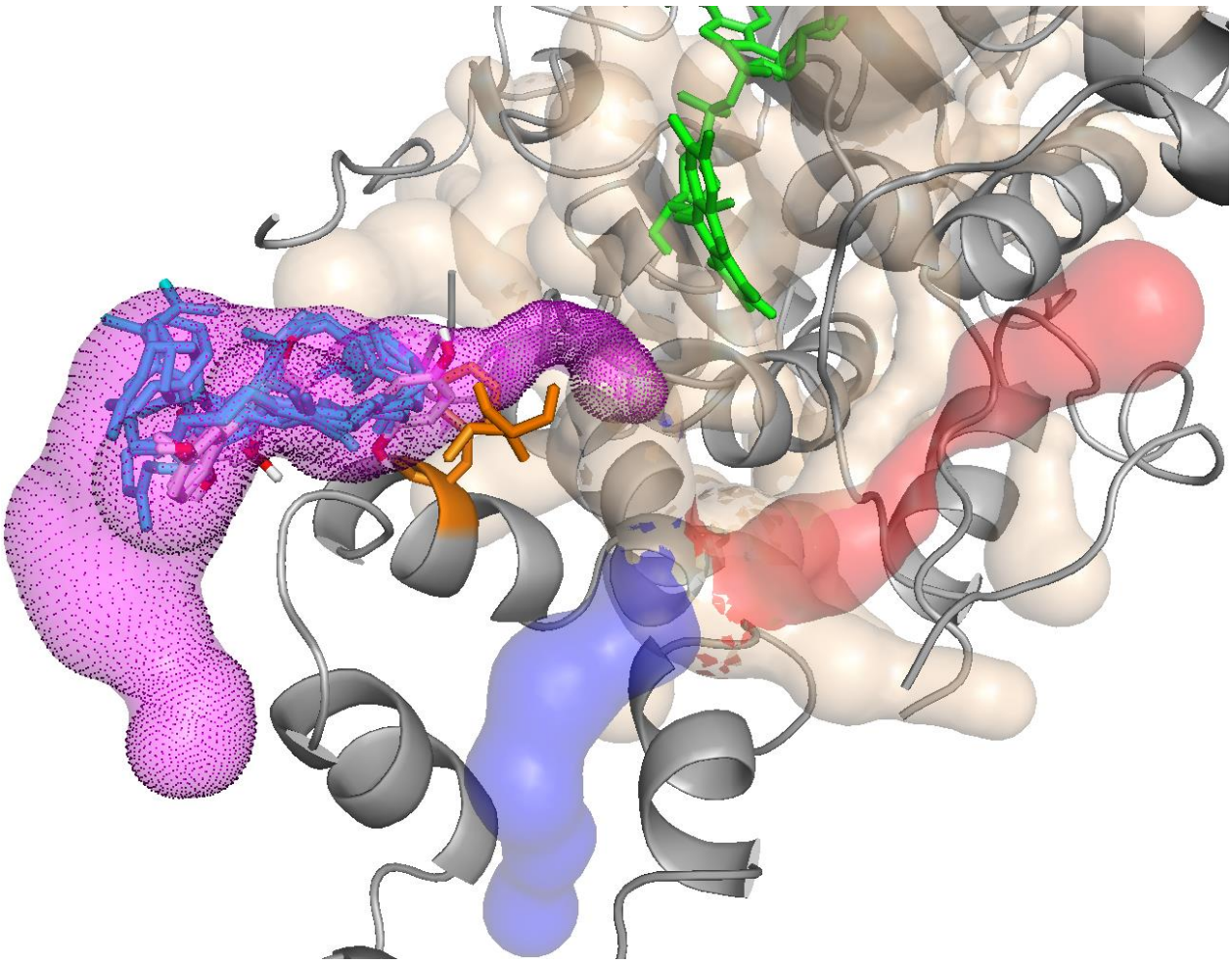

FIG 4.26A: *Docking results for 4-HP6 into the I-TASSER Coq6 model re face tunnel 1 (purple volume) Ghost tunnels are shown in beige; re face tunnel 2 in blue, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.26A shows the poses resulting from docking of our model substrate to a frame selected for maximum diameter of the *re* face tunnel 1, represented in purple. As can be seen from the figure, this tunnel has a bottleneck (indicated by the residues in orange stick) too small to admit passage of the polyprenyl tail, despite it being partially occupied in one of the poses. This conformation of the I-TASSER Coq6 model shows a concomitant opening of the *si*-face tunnel (shown as the red volume) and while some conformations are found that can traverse it, none of them place the aromatic head in the active site. Indeed, as we can see from **Figure 4.26 panels B and C**, the intersection of the main tunnel system does not occur in front of the FAD isoalloxazine. This is a recurrent feature of the I-TASSER model's tunnel system, and it means that even if the tunnels were to be traversable, they substrate's aromatic head could never achieve a catalytically plausible position. This is a rather fundamental flaw in the I-TASSER based model, making it unlikely that docking into any conformation would produce plausible enzyme-substrate complexes.

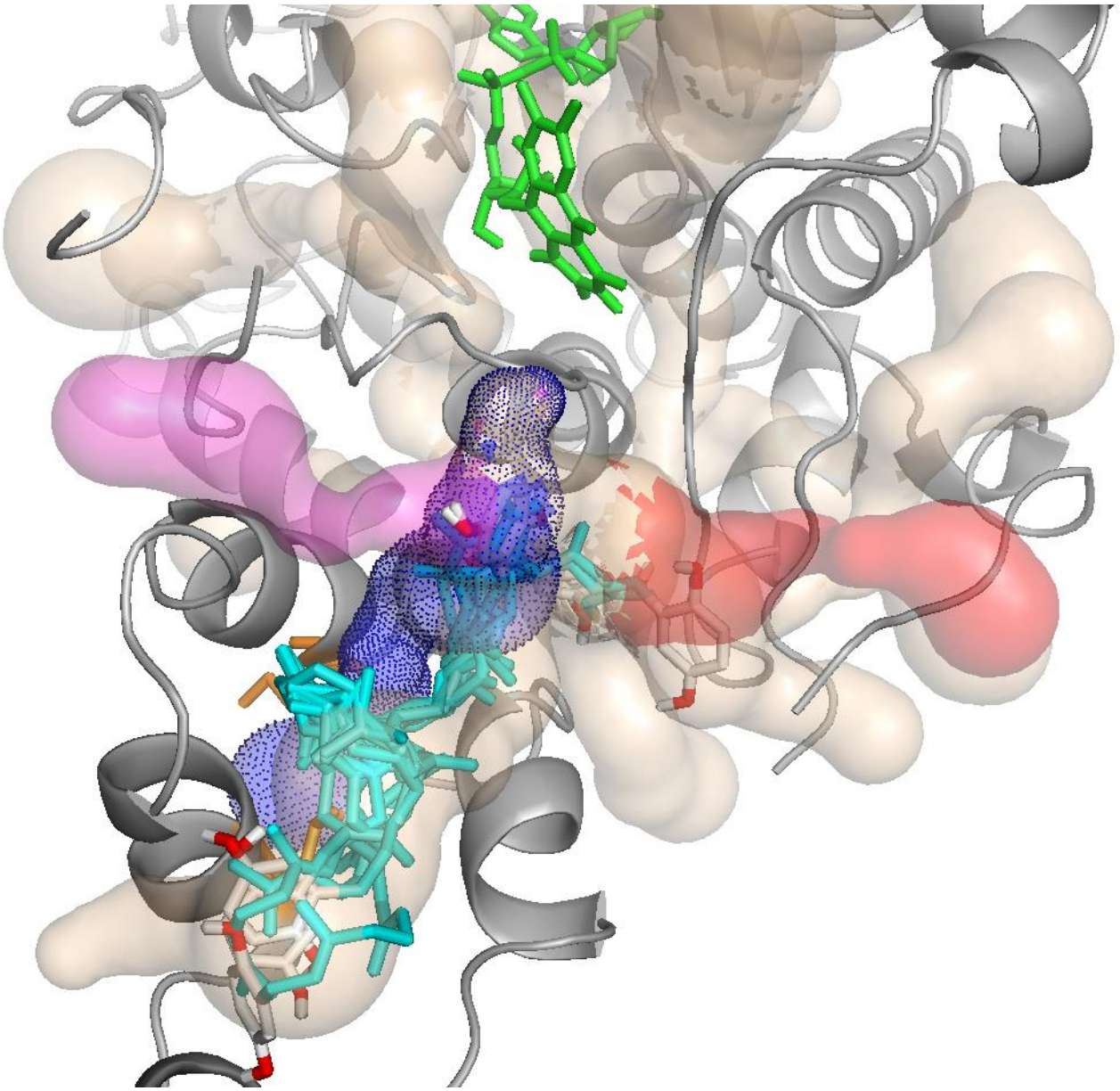

FIG 4.26B: *Docking results for 4-HP6 into the I-TASSER Coq6 model re face tunnel 2 (blue volume) Ghost tunnels are shown in beige; re face tunnel 1 in purple, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.26B shows the substrate poses resulting from docking to a receptor conformation corresponding to a maximum in the *re* face 2 tunnel bottleneck. The helix bundle conformation of the I-TASSER model's C-terminus does not permit a bottleneck large enough to allow substrate transit. Again, in this frame we see that the I-TASSER model's tunnel system does not converge to a volume in front of the FAD isoalloxazine.

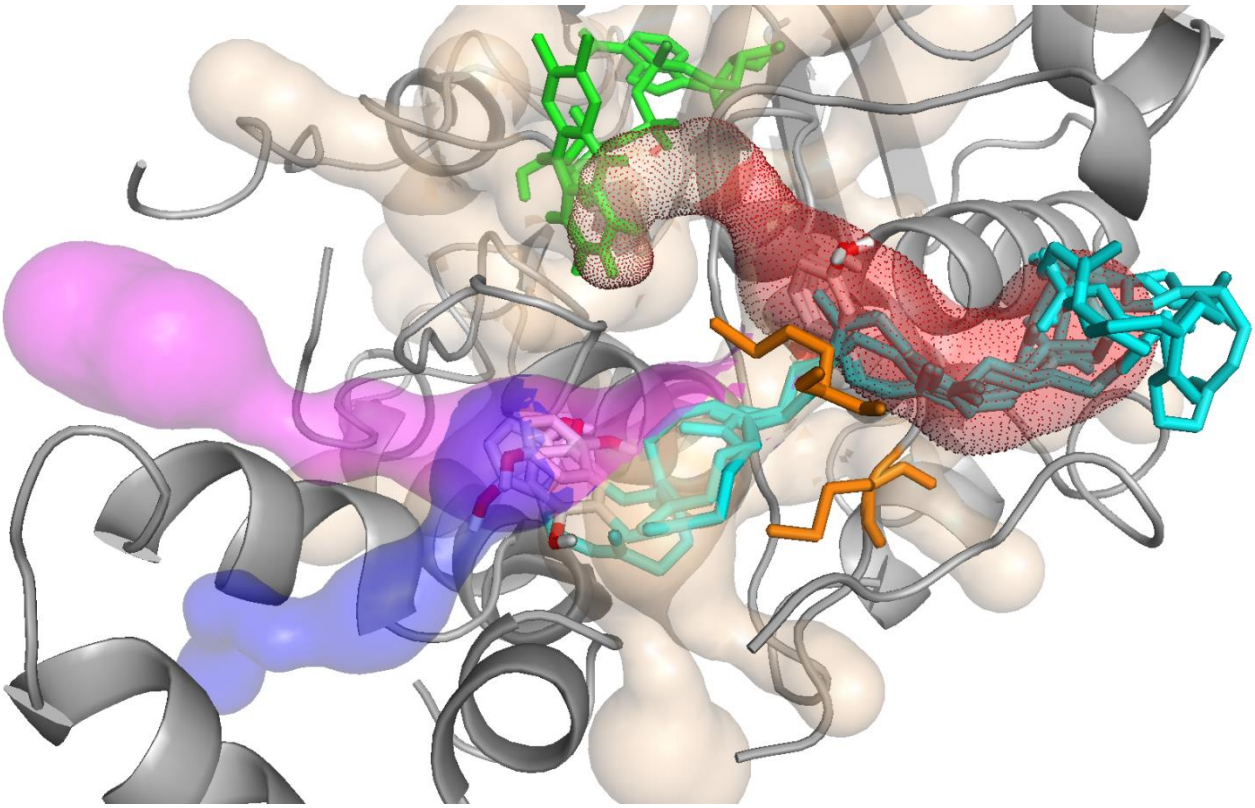

FIG 4.26C: *Docking results for 4-HP6 into the I-TASSER Coq6 model si face tunnel 2 (red volume) Ghost tunnels are shown in beige; re face tunnel 1 in purple, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.26C shows a similar result. The *si* face tunnel is not large enough to admit passage of the polyprenyl tail. Instead, the polyprenyl tail can only be accommodated by a larger cavity further in, but this is not in front of the FAD isoalloxazine.

In conclusion, the I-TASSER model seems to contain a fundamental inability to place the substrate in a catalytic pose because of the relative position of the tunnel system and the FAD.

5.1.2 Substrate docking into the ROBETTA Coq6 model

The results from docking our model substrate into the ROBETTA model are presented below in **Figure 27**.

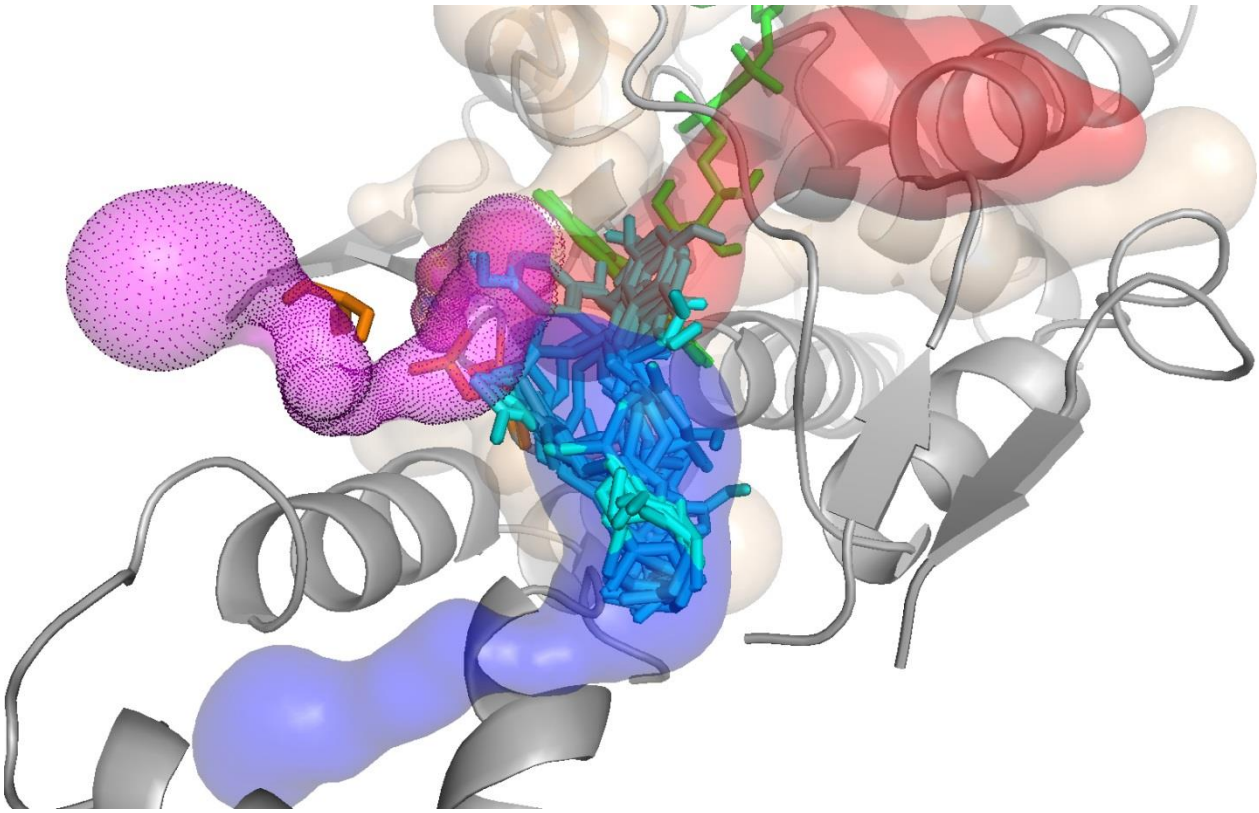

FIG 4.27A: *Docking results for 4-HP6 into the ROBETTA Coq6 model re face tunnel 1 (purple volume) Ghost tunnels are shown in beige; re face tunnel 2 in blue, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.27A shows the poses resulting from docking to a frame selected for maximum diameter of the *re* face tunnel 1 (in purple). As can be seen in the figure, this tunnel has a bottleneck (at the site of the residues in orange stick) too small to admit passage of the polyprenyl tail, disallowing even partial occupancy.

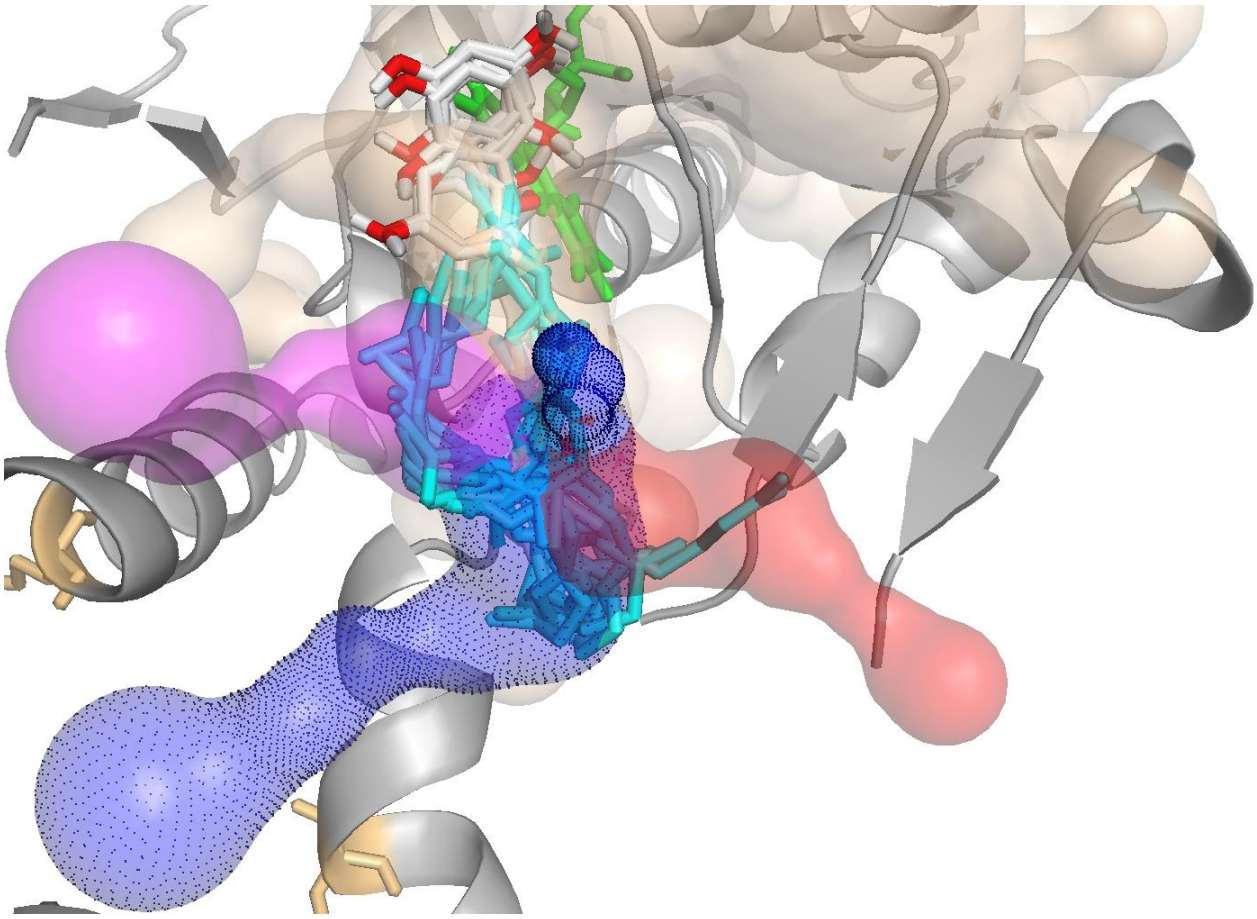

FIG 4.27B: *Docking results for 4-HP6 into the ROBETTA Coq6 model re face tunnel 2 (blue volume) Ghost tunnels are shown in beige; re face tunnel 1 in purple, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.27B shows the poses resulting from docking into a receptor conformation corresponding to a maximum in the *re*-face tunnel 2 diameter. The *re*-face tunnel collapses to a small diameter which cannot allow passage of the tail, forcing the substrate to assume a highly folded conformation inside the protein.

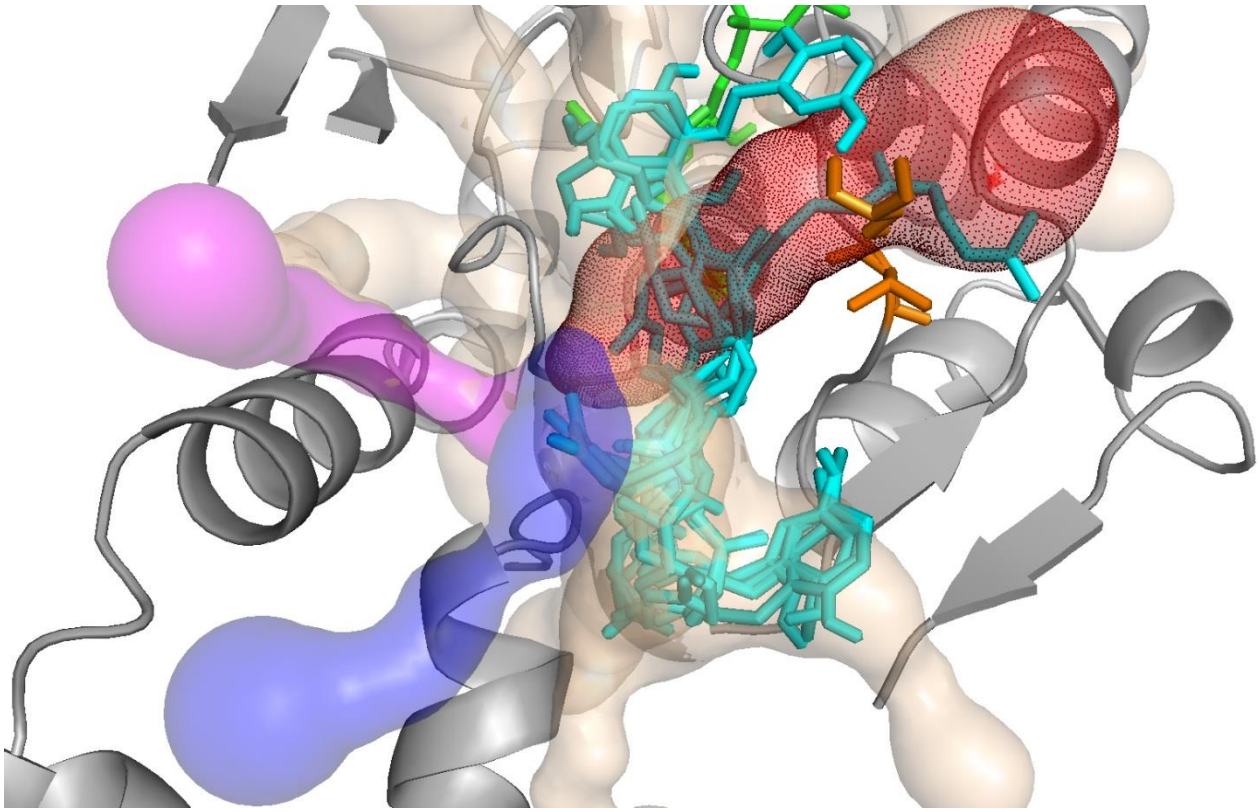

FIG 4.27C: *Docking results for 4-HP6 into the ROBETTA Coq6 model si face tunnel 2 (red volume) Ghost tunnels are shown in beige; re face tunnel 1 in purple, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.27C shows docking into a conformation of maximum *si* face tunnel diameter. Although one pose is found where the tail can traverse this tunnel, it does not place the aromatic head in front of the isoalloxazine.

While one of the tunnels in the ROBETTA model are traversable, this model shows a more realistic geometry in the sense that it can place the intersection of the tunnel system in front of the FAD isoalloxazine.

5.1.3 Substrate docking into the RATIONAL Coq6 model

The results from docking our model substrate into the RATIONAL model are presented below in **Figure 4.28**.

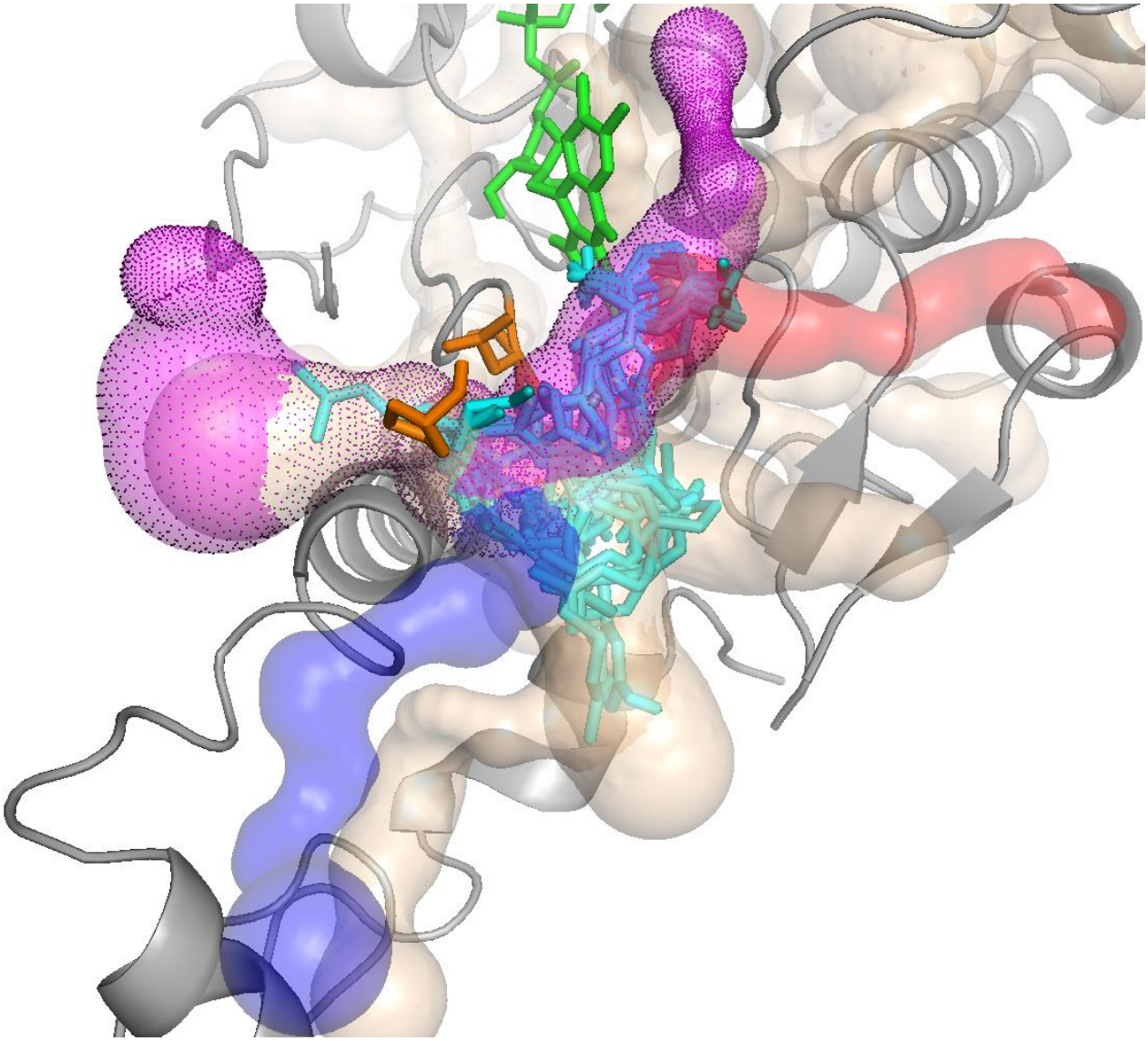

FIG 4.28A: *Docking results for 4-HP6 into the RATIONAL Coq6 model re face tunnel 1 (purple volume) Ghost tunnels are shown in beige; re face tunnel 2 in blue, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.28A shows the poses resulting from docking into a frame selected for a maximum diameter of *re* face tunnel 1. This job finds conformations which allow passage of the polyprenyl tail, extending from the surface to the active site (directly in front of the FAD isoalloxazine) passing through a bottleneck formed by L382 and P249 (shown in orange stick in **Figure 28A**). It also shows conformations where the polyprenyl tail can be completely folded into the active site.

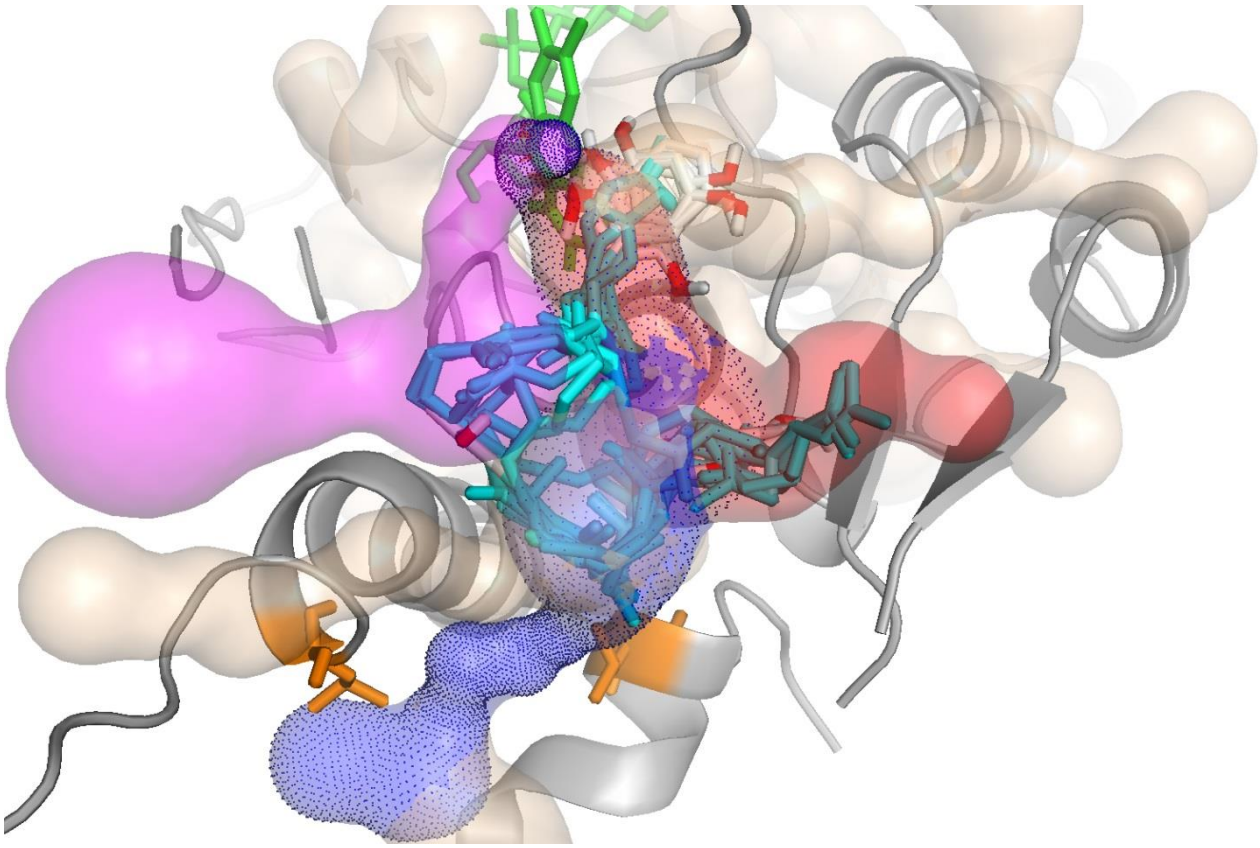

FIG 4.28B: *Docking results for 4-HP6 into the RATIONAL Coq6 model re face tunnel 2 (blue volume) Ghost tunnels are shown in beige; re face tunnel 1 in purple, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.28B shows docking into a conformation with a maximum in the diameter of *re* face tunnel 2, revealing that it is not traversable.

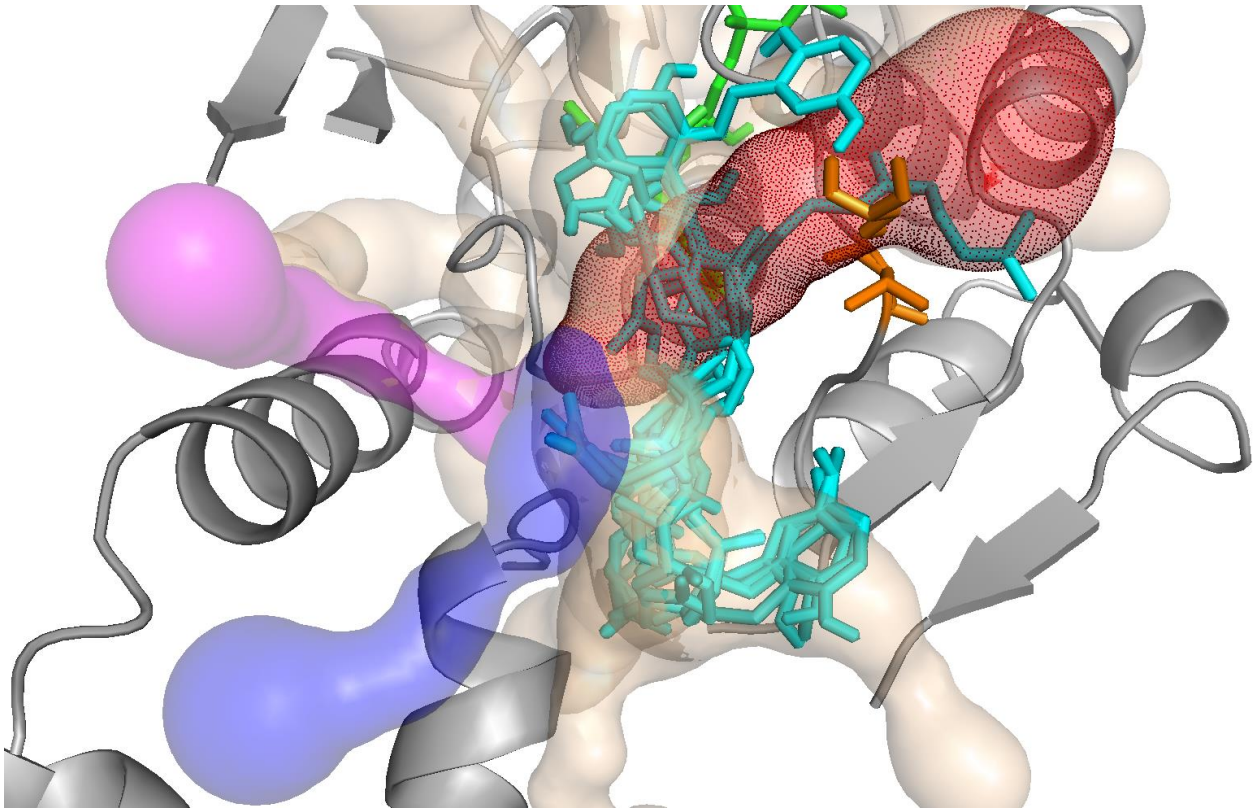

FIG 4.28C: *Docking results for 4-HP6 into the RATIONAL Coq6 model si face tunnel 2 (red volume) Ghost tunnels are shown in beige; re face tunnel 1 in purple, and si face tunnel in red. The FAD is represented in green, the substrate is represented in cyan sticks, and the substrate bottleneck residues in orange.*

Figure 4.28C shows the results of docking into the *si* face tunnel, which reveals that it is not traversable either.

5.1.4 Conclusion of Round 1 of substrate docking

Taken together, these results show that only *re* face tunnel 1 of the RATIONAL model is passable by the substrate and allows placement of the aromatic head in front of the FAD isoalloxazine. In addition, this study shows an interesting relationship between the tunnel systems and the active site geometry. This is an outstanding feature of the I-TASSER model, wherein the tunnel system does not converge to a point in front of the FAD, making the model generally unsuitable for generating enzyme-substrate complexes.

5.2 Round 2 of docking: the RATIONAL model and an active site geometry descriptor

The goal of this modeling process is to identify specific residues in Coq6 essential to enzyme-substrate interactions. According to our analysis thus far, any such residues are likely to form part of a substrate access tunnel, which we propose is *re* face tunnel 1, which forms a tunnel traversable by a model substrate. The Coq6 conformations corresponding to this state also allows placement of the aromatic head in front of the FAD. At this stage, we would like to refine our docking results in order to have more

accurate predictions of enzyme-substrate interactions. Our first round of docking established a feature of gross anatomy required for substrate binding: a substrate access tunnel large enough to admit substrate passage. Now we will turn our attention to the detailed anatomy of the active site, to see if the docking procedure can find conformations which are catalytically plausible in the context of the knowledge of the enzymatically active PHBH enzyme-substrate coordinates. We will perform a second round of docking, focusing only on the RATIONAL model, since it is the only one with passable substrate access tunnel. In this second round we will apply additional geometric selection criteria describing the detailed conformation of the Coq6 RATIONAL model active site to frame selection from molecular dynamics trajectories.

Since we do not have experimental coordinates for Coq6 enzyme-substrate complexes, or any other Q biosynthesis enzymes, we will use the binding mode of pHB in PHBH as described in the 1PBE crystal structure.[20] Since this PDB structure is our main structural reference for a catalytically active enzyme-substrate complex in this type of enzyme, we will first perform a comparison of the PHBH active site to the RATIONAL Coq6 model active site.

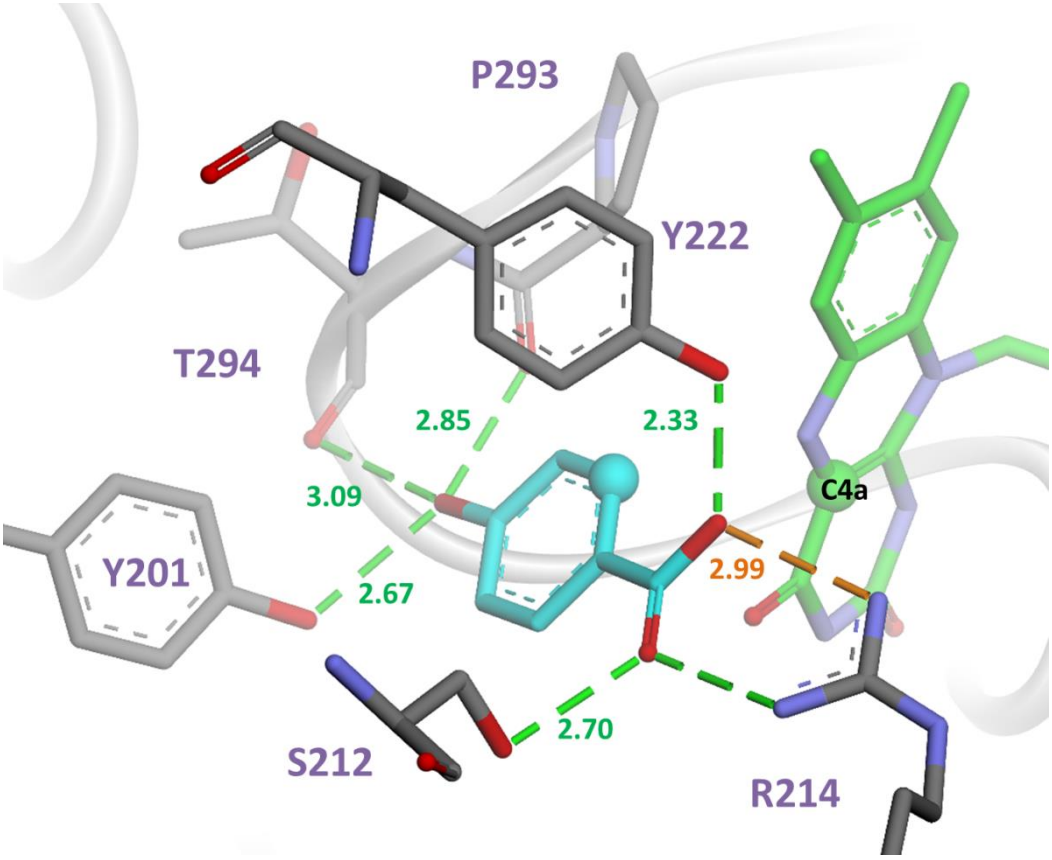

FIG 4.29: *The active site of PHBH structure 1PBE. The substrate, para-hydroxybenzoate, is shown in cyan stick with its hydroxylation target carbon shown as a cyan sphere. The FAD is shown in green, with its C4a carbon (which will bear the reactive peroxo group) shown as a green sphere. Residues binding the substrate through hydrogen bonds are shown in grey sticks. The substrate's hydroxyl group is hydrogen bonded to three residues (Y201, P293, T294), two of which (P293 and T294) bond through backbone oxygens. The substrate's carboxyl group forms hydrogen bonds with different residues (S212, R214, Y222). Most notable is R214, presenting the bidentate guanidinium group to bind the carboxyl oxygens.*

An inspection of the PHBH active site reveals that the substrate (as shown in **Figure 4.29**), a di-substituted aromatic ring similar to the aromatic head of many Q biosynthesis intermediates, is hydrogen bonded to six residues. The substrate's hydroxyl groups is hydrogen bonded to three residues: Y201, P293, and T294 at distances of 2.67Å, 2.85Å, and 3.09Å, respectively. The hydrogen bonds of P293 and T294 are made with their backbone oxygens; these positions are highly conserved in the sequences of Coq6 family enzymes (as shown in the Coq6 family multiple sequence alignment presented in Annex 2) and highly conserved in structure among crystallized enzymes of this global fold. The carboxyl group of the substrate is involved in four hydrogen bonds with residues S212, R214, and Y222. S212 and Y222 contribute one hydrogen bond each, while two are contributed by a bidentate interaction with the guanidinium group of R214. These hydrogen bonds serve to hold the substrate in a precise orientation and distance relative to the FAD isoalloxazine, as necessary for catalysis. The key point is that in the 1PBE structure of PHBH as complexed with FAD and substrate pHB, all molecules have complementary conformations for binding.

In particular, we are interested in the shape complementarity between enzyme and substrate. Assuming a homologous pattern of enzyme-substrate contacts, plausible given the similarity of their respective substrates, the coordinates of the 1PBE structure can tell us something very important: the shape of the active site that can form a catalytically competent complex. The precise shape of the PHBH active site can be defined explicitly in terms of interatomic distances that involve only the atoms of the enzyme and cofactor. That is to say, we can describe the shape of the active site when it is bound to substrate without needing to refer to coordinates of the substrate itself. Having a set of interatomic reference distances to define this shape also enables us to recognize when it occurs in an ensemble – such as the ensembles we have generated by molecular dynamics – even if the ensemble was generated without the substrate. This method gives us a specific tool to test the conformational selection hypothesis for the specific enzyme of interest: the enzyme-substrate geometry reference PHBH, and the enzyme of the present study, Coq6.

A receptor-based enzyme geometry scoring function was first developed to characterize the catalytically plausible enzyme-cofactor-substrate complex of PHBH in structure 1PBE. We recorded the interatomic distances from the PHBH 1PBE co-crystal structure corresponding to the enzyme-substrate hydrogen bonds. We also included the distance between the FAD C4a carbon (which will bear the reactive peroxo group) and the hydroxylation target carbon on the substrate. These sets of interatomic distances are shown in **Figure 4.30** below.

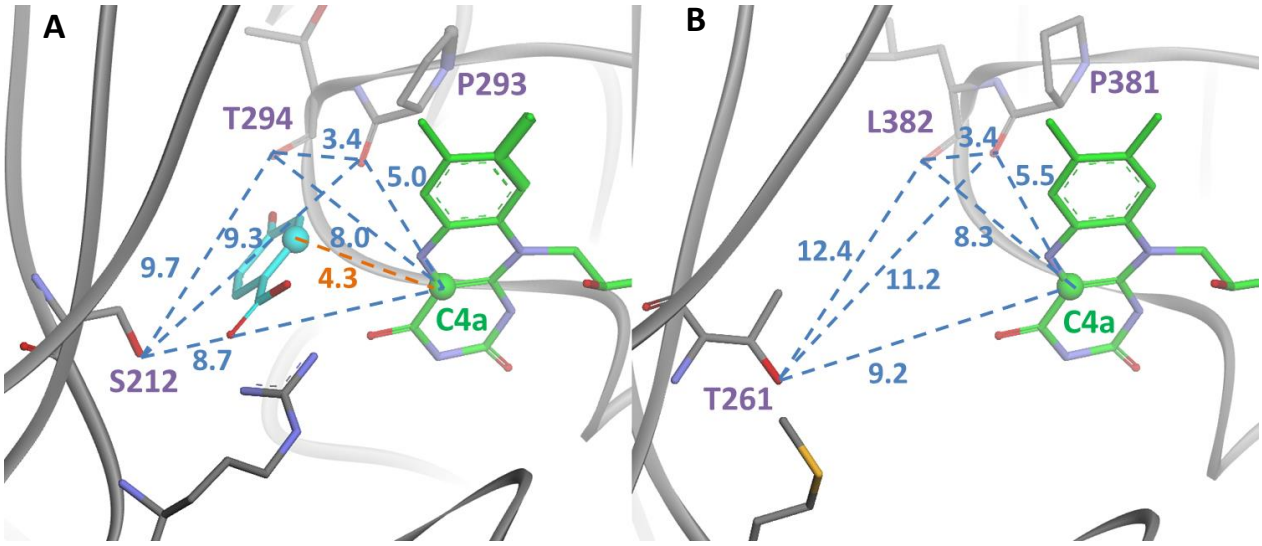

FIG 4.30: *A) The active site of PHBH structure 1PBE showing the receptor based interatomic distances used in creating the geometric descriptor and scoring function. B) The active site of the RATIONAL Coq6 model (before MD) showing the distances between homologous atoms. We note the absence of an equivalent to PHBH R214, used to engage the substrate's carboxyl group.*

These distances were used to create a scoring function, denoted S, allowing us to quantify the similarity between the crystal reference distance set in PHBH structure 1PBE and the homologous distance set from simulated conformations of Coq6. The scoring function S was constructed as the sum of the differences between the catalytic site interatomic distances measured in the 1PBE crystal structure and those sampled from MD simulations of the Coq6-FAD complex.

$$S = \Sigma \left[d_{ij}(\text{Coq6}) - d_{ij}(\text{1PBE}) \right]$$

This scoring function is applied to the enzyme coordinates from every frame of the MD trajectory, providing a criterion for selecting enzyme conformations compatible with catalytic substrate binding to be used in subsequent docking. A low score indicates a high structural similarity between the substrate bound PHBH active site in the 1PBE structure and the substrate free Coq6 conformations sampled during MD.

This protocol was first developed and tested on the PHBH-FAD complex described in the 1PBE PDB structure. The substrate was removed from the starting structure and molecular dynamics was performed and then analyzed with the scoring function. Re-docking of pHB into the PHBH-FAD complex was able to reproduce the crystal pose of the substrate. This initial result (documented in Annex 4) suggested that PHBH enzyme-substrate system may be work through conformational selection. Since the RATIONAL Coq6 model is structurally similar to PHBH, we posit that molecular dynamics trajectories of this Coq6 model may contain conformations compatible with substrate binding. Therefore, we identified homologous atoms in the Coq6 structure and performed the same analysis.

The strength of the geometric descriptor developed from PHBH is that it can be applied to the structurally homologous enzyme Coq6, which hydroxylates a structurally homologous substrate, by mapping it onto homologous atoms. A superposition of the Coq6 and PHBH structures allows us to

identify structurally homologous atoms as illustrated in **Figure 4.30**. Coq6 P381 and L382 are homologous to PHBH P293 and T294, where these residues hydrogen bond to the substrate via their backbone oxygens. Substrate contacts with the backbone are likely to be conserved in Coq6 because functional mutations to this sequence position do not change the structural position of the backbone. Coq6 T261 is structurally homologous to PHBH S212, and can be a homologous hydrogen bond donor to the Coq6 substrate. However, we note the absence of a Coq6 residue homologous to PHBH R214. A review of protein structures binding carboxylated substrates indicates this chemical group is almost always bound with an arginine's guanidinium group.[20] [21] [22] Therefore, the absence of a homologous residue in Coq6 suggests that the Coq6 substrate is not carboxylated. This also suggests that the active site of Coq6 is not structured to bind a carboxylated substrate, but rather, the di-hydroxylated species 3-hexaprenyl-4-hydroxyphenol (4-HP6). This is consistent with experimental results, in which structurally intact but catalytically inactive Coq6 mutants accumulate this same molecule. In addition, such mutants cultured in media to depend on 4-aminobenzoate accumulate 3-hexaprenyl-4-aminophenol, the C4-aminated equivalent of 4-HP6.[23] Therefore, we will use 4-HP6 as a model substrate in our docking studies.

Using the mapping of homologous atoms shown in **Figure 4.30** we will maintain the reference distances from PHBH, but now we will compare them to the homologous distances over the molecular dynamics trajectory of the Coq6 model. Below is a figure showing the behavior of the scoring function over the course of the Coq6 simulation.

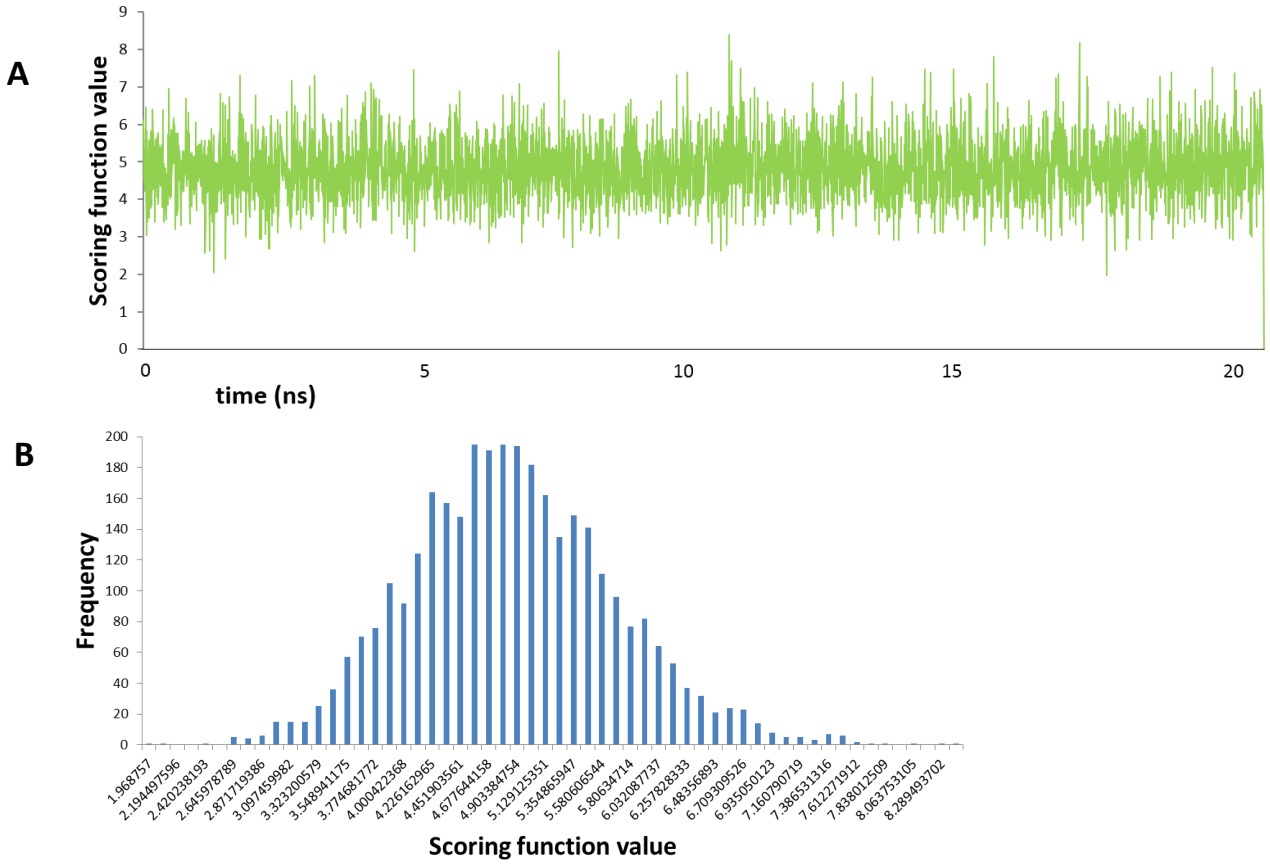

FIG 4.31: *A) A plot of the active site descriptor scoring function over the 20 ns trajectory of the RATIONAL Coq6 model. Low scoring conformations from the latter half of the simulation were selected for the*

second round of substrate docking. B) A histogram of the scoring function suggests a unimodal distribution

The scoring function allows us to select the simulation frames most likely to contain Coq6 conformations compatible with binding an aromatic substrate in a catalytically plausible position.

We now have two explicit geometric descriptions of the essential functional regions of Coq6: the substrate access channel and the active site. This gives us two powerful filters that we can apply in a step-wise process for finding relevant conformations in our simulated ensembles of Coq6 substrate-free conformations. Our first geometric condition for catalytically plausible Coq6 conformations is that the substrate access channel be traversable: this is described by the bottleneck diameter metric. The second geometric condition for catalytically plausible Coq6 conformation is that the active site must be in a conformation that can allow the binding of the substrate in a manner similar to the substrate in PHBH. Therefore, in our conformational filtering procedure, we consider only the second half of the simulation trajectories (the last 10 ns) where the protein structure has stabilized. We first calculate the maximum diameter of the narrowest points for each frame of the simulation, and then re-sort the frames by this value. We make a first selection of the top half of the frames in this ranking. We then sort these remaining frames by the scoring function S, based on the geometric descriptor of the Coq6 active site. This gives us a ranked list of conformations that have both an open substrate access channel and an active site geometry compatible with catalysis. Finally, we select the top ranked frames from this list and use them for substrate docking. Applying both selection criteria of enzyme geometry to the trajectory enables us to select a better Coq6 conformation for substrate docking. An example of such a conformation and the resulting docked poses are shown below.

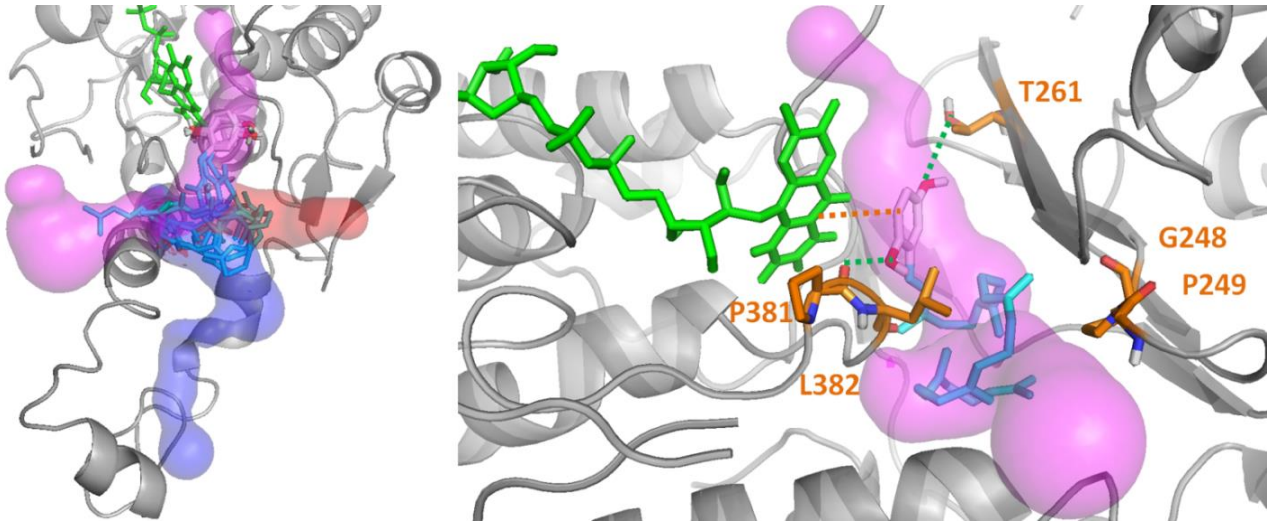

FIG 4.32: *Resulting poses for docking of 4-HP6 into a RATIONAL Coq6 model conformation selected using the two criteria of access channel diameter and active site geometry score. A) The top 10 substrate poses (in cyan stick). B) The docked substrate pose showing the best FAD C4a – substrate C5 distance: 4.68 Å. This compares favorably to the homologous distance of 4.32 Å in the PHBH 1PBE crystal structure. Active site substrate hydrogen bonding residues P381 and T261 are shown in orange stick, as are re face tunnel 1 bottleneck residues L382 and P249.*

The results of this second round of docking show a much more consistent positioning of the aromatic head in the active site. **Indeed, the main variation is in the conformation of the hexaprenyl tail, which**

can either traverse the tunnel or curl up entirely inside the active site volume. The aromatic head hydroxyl groups consistently form hydrogen bonds with the backbone oxygens of P381 (at a distance of 2.49Å) and T261 (at a distance of 2.38Å). The hexaprenyl tail forms contacts with the *re* face tunnel 1 bottleneck residues L382 and P249.

These four residues are highly conserved as computed in the ConSurf MSA as well. Indeed, the residues lining the *re* face channel are all highly conserved, as shown in Figure 4.33 below.

residue number	residue type	ConSurf conservation score
85	VAL	8
86	SER	9
114	LEU	7
222	ALA	7
226	VAL	9
242	GLN	9
244	PHE	9
247	THR	8
248	GLY	9
249	PRO	9
251	ALA	9
253	LEU	9
261	THR	8
263	VAL	9
264	TRP	9
265	SER	9
354	PHE	9
356	LEU	9
380	HIS	9
382	LEU	9
383	ALA	9
384	GLY	9
439	PHE	9
440	LYS	9
443	HIS	7
444	THR	7
480	FAD	9
246	PRO	8
250	ILE	7
266	SER	8
381	PRO	9
438	LEU	8

FIG 4.33: *Residue conservation (calculated by ConSurf) for the re face tunnel 1 (calculated by CAVER) of the RATIONAL Coq6 model. The residue conservation column is also colored according to the ConSurf color scheme.*

6. Conclusion

We have produced three homology models of the Coq6 enzyme, one based on a manually curated alignment of multiple rationally selected templates, and two through the use of automated servers: I-TASSER and ROBETTA. While these models showed structural stability after 20ns of molecular dynamics, calculations of accessible volumes show differences in their possible substrate access tunnels. Accessible volume calculations reveal three types of tunnels leading from the protein surface and converging to the active site. In order to explore the ability of each type of tunnel in each model to allow substrate access to the active site, we characterized the diameter of each tunnel at its narrowest point through the definition of pairs of bottleneck residues. We selected Coq6 model conformations from dynamics displaying maxima in channel diameters for all three models for subsequent substrate docking in order to assess their traversability by a model substrate, 3-hexaprenyl-4-hydroxyphenol. This model substrate was selected after a preliminary molecular docking screening to determine the optimal polyprenyl tail length for more detailed docking studies. This model substrate choice is also consistent with the best experimental data on the identity of the Coq6 substrate, which is substrate accumulation in enzymatically inactive Coq6 yeast strains.[23]

The I-TASSER model shows a tunnel system that does not converge in front of the FAD isoalloxazine, making it unlikely that any docking attempts will place a substrate in a catalytically plausible position. In addition, the tunnels of this model are consistently too narrow to permit passage of the substrate. The ROBETTA model shows better positioning of its tunnel system relative to the FAD, but its tunnels are too narrow as well. **This analysis reveals that only the *re* face tunnel 1 of the RATIONAL model can permit the aromatic head of the substrate to reach the active site, making it the only model we will continue to study.**

Despite the experimental ambiguity in the mapping of enzymes to substrates, the development of our theoretically based RATIONAL Coq6 model can suggest some substrates as being more likely than others for this enzyme. This work proposes 3-hexaprenyl-4-hydroxyphenol as the Coq6 substrate. By comparison to the active site of a functional homolog co-crystallized with a similar substrate, PHBH, we derived a set of receptor-based interatomic distances to describe the substrate bound active site conformation. This set of distances was used to score and select conformations from molecular dynamics trajectories. These conformations, selected for maximum geometric similarity to the PHBH structure, were used for substrate docking. **The results of this docking confirm the importance of the bottleneck residues identified for the functional: P249 and L382**. The docking also identifies two residues likely to form hydrogen bonds with the substrate's hydroxyl groups, P381 and T261. **Substrate docking was a first *in silico* test of the functionality of *re* face tunnel 1. The next step is to test this tunnel experimentally through the creation of site directed mutations designed to block the tunnel.**

References

[1] Vogt, Austin D., and Enrico Di Cera. "Conformational Selection or Induced-Fit? A Critical Appraisal of the Kinetic Mechanism." *Biochemistry* 51, no. 30 (July 31, 2012): 5894–5902. doi:10.1021/bi3006913.

[2] Huang, Sheng-You, and Xiaoqin Zou. "Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking." *Proteins* 66, no. 2 (February 1, 2007): 399–421. doi:10.1002/prot.21214.

[3] Trott, Oleg, and Arthur J. Olson. "AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading." *Journal of Computational Chemistry* 31, no. 2 (January 30, 2010): 455–61. doi:10.1002/jcc.21334.

[4] Gatti, D. L., B. A. Palfey, M. S. Lah, B. Entsch, V. Massey, D. P. Ballou, and M. L. Ludwig. "The Mobile Flavin of 4-OH Benzoate Hydroxylase." *Science (New York, N.Y.)* 266, no. 5182 (October 7, 1994): 110–14.

[5] Lah, M. S., B. A. Palfey, H. A. Schreuder, and M. L. Ludwig. "Crystal Structures of Mutant Pseudomonas Aeruginosa P-Hydroxybenzoate Hydroxylases: The Tyr201Phe, Tyr385Phe, and Asn300Asp Variants." *Biochemistry* 33, no. 6 (February 15, 1994): 1555–64.

[6] Ortiz-Maldonado, M., D. Gatti, D. P. Ballou, and V. Massey. "Structure-Function Correlations of the Reaction of Reduced Nicotinamide Analogues with P-Hydroxybenzoate Hydroxylase Substituted with a Series of 8-Substituted Flavins." *Biochemistry* 38, no. 50 (December 14, 1999): 16636–47.

[7] Wang, Jian, Mariliz Ortiz-Maldonado, Barrie Entsch, Vincent Massey, David Ballou, and Domenico L. Gatti. "Protein and Ligand Dynamics in 4-Hydroxybenzoate Hydroxylase." *Proceedings of the National Academy of Sciences* 99, no. 2 (January 22, 2002): 608–13. doi:10.1073/pnas.022640199.

[8] Ortiz-Maldonado, M., S. M. Aeschliman, D. P. Ballou, and V. Massey. "Synergistic Interactions of Multiple Mutations on Catalysis during the Hydroxylation Reaction of P-Hydroxybenzoate Hydroxylase: Studies of the Lys297Met, Asn300Asp, and Tyr385Phe Mutants Reconstituted with 8-Cl-Flavin." *Biochemistry* 40, no. 30 (July 31, 2001): 8705–16.

[9] Ortiz-Maldonado, Mariliz, Lindsay J. Cole, Sara M. Dumas, Barrie Entsch, and David P. Ballou. "Increased Positive Electrostatic Potential in P-Hydroxybenzoate Hydroxylase Accelerates Hydroxylation but Slows Turnover." *Biochemistry* 43, no. 6 (February 17, 2004): 1569–79. doi:10.1021/bi030193d.

[10] Cole, Lindsay J., Barrie Entsch, Mariliz Ortiz-Maldonado, and David P. Ballou. "Properties of P-Hydroxybenzoate Hydroxylase When Stabilized in Its Open Conformation." *Biochemistry* 44, no. 45 (November 15, 2005): 14807–17. doi:10.1021/bi0512142.

[11] Ortiz-Maldonado, Mariliz, Barrie Entsch, and David P. Ballou. "Oxygen Reactions in P-Hydroxybenzoate Hydroxylase Utilize the H-Bond Network during Catalysis." *Biochemistry* 43, no. 48 (December 7, 2004): 15246–57. doi:10.1021/bi048115t.

[12] Entsch, Barrie, Lindsay J. Cole, and David P. Ballou. "Protein Dynamics and Electrostatics in the Function of P-Hydroxybenzoate Hydroxylase." *Archives of Biochemistry and Biophysics* 433, no. 1 (January 1, 2005): 297–311. doi:10.1016/j.abb.2004.09.029.

[13] Ballou, David P., Barrie Entsch, and Lindsay J. Cole. "Dynamics Involved in Catalysis by Single-Component and Two-Component Flavin-Dependent Aromatic Hydroxylases." *Biochemical and Biophysical Research Communications* 338, no. 1 (December 9, 2005): 590–98. doi:10.1016/j.bbrc.2005.09.081.

[14] Frederick, K. K., D. P. Ballou, and B. A. Palfey. "Protein Dynamics Control Proton Transfers to the Substrate on the His72Asn Mutant of P-Hydroxybenzoate Hydroxylase." *Biochemistry* 40, no. 13 (April 3, 2001): 3891–99.

[15] Celniker, Gershon, Guy Nimrod, Haim Ashkenazy, Fabian Glaser, Eric Martz, Itay Mayrose, Tal Pupko, and Nir Ben-Tal. "ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function." *Israel Journal of Chemistry* 53, no. 3–4 (April 1, 2013): 199–206. doi:10.1002/ijch.201200096.

[16] Chovancova, Eva, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, et al. "CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures." *Plos Computational Biology* 8, no. 10 (October 2012): e1002708. doi:10.1371/journal.pcbi.1002708.

[17] Clarke, Catherine F. "New Advances in Coenzyme Q Biosynthesis." *Protoplasma* 213, no. 3–4 (September 2000): 134–47. doi:10.1007/BF01282151.

[18] He, Cuiwen H., Letian X. Xie, Christopher M. Allan, UyenPhuong C. Tran, and Catherine F. Clarke. "Coenzyme Q Supplementation or over-Expression of the Yeast Coq8 Putative Kinase Stabilizes Multi-Subunit Coq Polypeptide Complexes in Yeast Coq Null Mutants." *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1841, no. 4 (April 2014): 630–44. doi:10.1016/j.bbalip.2013.12.017.

[19] Humphrey, William, Andrew Dalke, and Klaus Schulten. "VMD: Visual Molecular Dynamics." *Journal of Molecular Graphics* 14, no. 1 (February 1996): 33–38. doi:10.1016/0263-7855(96)00018-5.

[20] Schreuder, Herman A., Peter A. J. Prick, Rik K. Wierenga, Gerrit Vriend, Keith S. Wilson, Wim G. J. Hol, and Jan Drenth. "Crystal Structure of the P-Hydroxybenzoate Hydroxylase-Substrate Complex Refined at 1.9 Å Resolution: Analysis of the Enzyme-Substrate and Enzyme-Product Complexes." *Journal of Molecular Biology* 208, no. 4 (August 20, 1989): 679–96. doi:10.1016/0022-2836(89)90158-7.

[21] Gallagher, D. T., M. Mayhew, M. J. Holden, A. Howard, K. J. Kim, and V. L. Vilker. "The Crystal Structure of Chorismate Lyase Shows a New Fold and a Tightly Retained Product." *Proteins* 44, no. 3 (August 15, 2001): 304–11.

[22] Thoden, James B., Zhihao Zhuang, Debra Dunaway-Mariano, and Hazel M. Holden. "The Structure of 4-Hydroxybenzoyl-CoA Thioesterase from Arthrobacter Sp. Strain SU." *The Journal of Biological Chemistry* 278, no. 44 (October 31, 2003): 43709–16. doi:10.1074/jbc.M308198200.

[23] Ozeir, Mohammad, Ulrich Mühlenhoff, Holger Webert, Roland Lill, Marc Fontecave, and Fabien Pierrel. "Coenzyme Q Biosynthesis: Coq6 Is Required for the C5-Hydroxylation Reaction and Substrate Analogs Rescue Coq6 Deficiency." *Chemistry & Biology* 18, no. 9 (September 23, 2011): 1134–42. doi:10.1016/j.chembiol.2011.07.008.

Chapter 5 Testing the hypothesis of a Coq6 substrate access channel

1. Introduction

Thus far we have constructed a stable model of the wild type Coq6 enzyme and identified an evolutionarily conserved putative substrate access channel. We have performed an *in silico* test of the channel by docking a model substrate into it and have found that it can admit the substrate in a catalytically plausible pose. A straightforward approach to test this channel would be to introduce a site directed mutation in order to block the channel. Since our preceding analysis has already identified a bottleneck in this channel, we also have good starting points for proposing specific mutations. However, we would also like to identify specific active site residues which could be important for substrate binding. Therefore, we will refine our docking procedure by being more detailed and selective in our sampling of enzyme conformations from molecular dynamics.

Before continuing on this path we will review the literature for naturally occurring Coq6 mutations to see if our model is also consistent with these pre-existing mutant characterizations.

2. Review of known Coq6 mutants

Review of the literature reveals several mutations to Coq6 in human patients with deleterious effects. These are either point mutations or truncations, and are listed in **Table 5.1** below.

TABLE 5.1: *Human Coq6 mutations documented in the work of Heeringa 2011, Ozeir 2011, Zhang 2014, and Doimo 2014.*

Human Coq6 mutations	Source	Yeast Coq6 homologs mapped in this work
G255R	Heeringa 2011[1]	G248R
A353D	Heeringa 2011	A361D
R162X	Heeringa 2011	K155X
W188X	Heeringa 2011	W181X
W447X	Heeringa 2011	F455X
Q461fsX478	Heeringa 2011	M469X
D208H	Zhang 2014[2]	D201H
Y412C	Doimo 2014[3]	F420C
	Ozeir 2011[4]	G202V
	Ozeir 2011	G386A
	Ozeir 2011	N388D

The human Coq6 mutations are clinically known to result in a decrease of Coq6 function. Since we would like to develop a molecular understanding of how a number of these mutations affect Coq6 function, we used our alignment between *H. sapiens* Coq6 and *S. cerevisiae* Coq6 to identify the homologous residues in our model.

We manually aligned the human *H. sapiens* Coq6 sequence onto the *S. cerevisiae* Coq6 multiple sequence alignment computed by ConSurf (presented in Annex 2) and the homologous mutations are highlighted in **Figure 5.1** below.

```
      1          11         21           25        33          43
yCoq6 MFFSKVMLTR RILVRGLATA KSSA------ --PKLTDVLI VGGGPAGLTL AASIKNSPQL
hCoq6 MAARLVSRCG AVRAAPHSGP LVSWRRWSGA STDTVYDVVV SGGGLVGAAM ACALGYDIHF
      1          11         21          31          41          51

      53         63         73          83          93            103
yCoq6 KDLKTTLVDM VDLKDKLSDF YNSPPDYFTN RIVSVTPRSI HFLENNAGAT --LMHDRIQS
hCoq6 HDKKILLLEA GPKK-VLEKL S----ETYSN RVSSISPGSA TLLS-SFGAW DHICNMRYRA
      61         71         80          86          96          105

      111        121          129       137         147         157
yCoq6 YDGLYVTDGC SKATLDLA-- --RDSMLCMI EIINIQASLY NRISQYDSKK DSIDIIDNTK
hCoq6 FRRMQVWDAC SEALIMFDKD NLDD-MGYIV ENDVIMHALT KQLEAVSDR- --VTVLYRSK
      115        125        135         144         154           163

      167        177        181         191         201         211
yCoq6 VVNIKHSDPN D------PLS WPLVTLSNGE VYKTRLLVGA DGFNSPTRRF SQIPSRGWMY
hCoq6 AIRYTWP--- CPFPMADSSP WVHITLGDGS TFQTKLLIGA DGHNSGVRQA VGIQNVSWNY
      171        178        188         198         208         218

      221        231        240         250         260         270
yCoq6 NAYGVVASMK LEY-PPFKLR GWQRFLPTGP IAHLPMPENN ATLVWSSSER LSRLLLSLPP
hCoq6 DQSAVVATLH LS-EATENNV AWQRFLPSGP IALLPLSDTL SSLVWSTSHE HAAELVSMDE
      228        238        247         257         267         277

      280        290        300         310         320         330
yCoq6 ESFTALINAA FVLEDADMNY YYRTLEDGSM DTDKLIEDIK FRTEEIYATL KDESDIDEIY
hCoq6 EKFVDAVNSA FWSDADHTDF IDTAG----- ---AMLQYAV SLLKPTKVS- ------ARQL
      287        297        307            312     319               328

      340        350        360         370         380         390
yCoq6 PPRVVSIIDK TRARFPLKLT HADRYCTDRV ALVGDAAHTT HPLAGQGLNM GQTDVHGLVY
hCoq6 PPSVARVDAK SRVLFPLGLG HAAEYVRPRV ALIGDAAHRV HPLAGQGVNM GFGDISSLAH
      332        342        352         362         372         382

      400        410        420         430         440         450
yCoq6 ALEKAMERGL DIGSSLSLEP FWAERYPSNN VLLGMADKLF KLYHTNFPPV VALRTFGLNL
hCoq6 HLSTAAFNGK DLGSVSHLTG YETERQRHNT ALLAATDLLK RLYSTSASPL VLLRTWGLQA
      392        402        412         422         432         442
```

FIG 5.1: *Sequence alignment of H. sapiens Coq6 and S. cerevisiae Coq6. Manually curated in the context of the multiple sequence alignment of Coq6 family members identified by the ConSurf analysis (presented in Annex 2). Point mutations from human clinical literature are highlighted in green; truncation mutations from human clinical literature are highlighted in red; point mutants created only in yeast are highlighted in cyan.*

This allows us to map the location of the human Coq6 mutations onto our RATIONAL model of the yeast Coq6, as shown in **Figure 5.2** below.

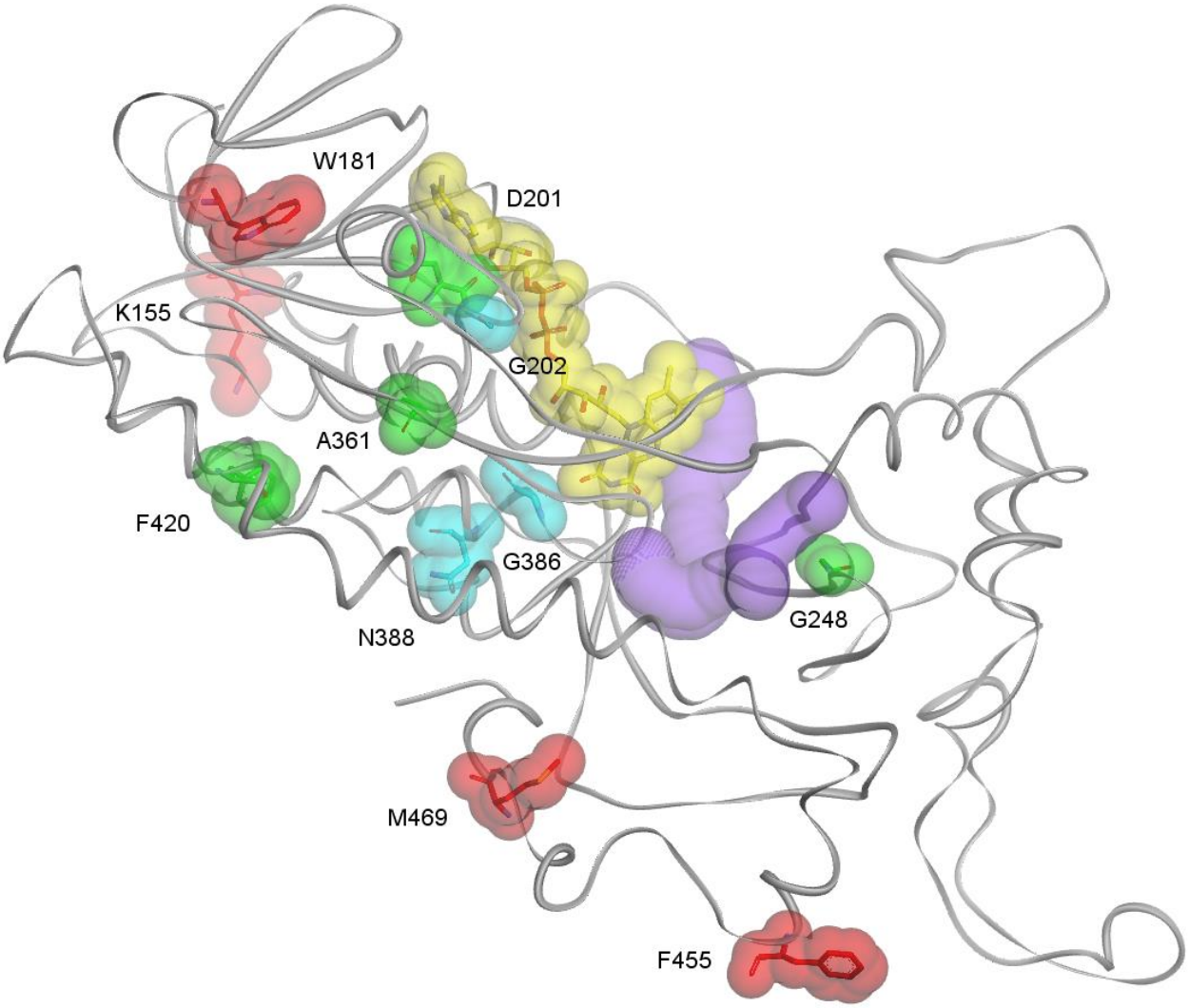

FIG 5.2: *Human and yeast Coq6 mutations documented in the literature as described in Table 5.1 mapped onto the RATIONAL yeast Coq6 model. The wild-type residues at the mutated positions are shown as sphere. FAD is shown in yellow, and the re face tunnel 1 is shown as the purple volume. Yeast Coq6 G248R, homologous to G255, is positioned near the entrance of the tunnel. Inactivating mutations mapped from the human Coq6 sequence are shown in green sphere; truncation mutation sites mapped from human Coq6 are shown in red sphere; inactivating point mutations created specifically for yeast Coq6 are shown in cyan sphere.*

The truncation mutations, illustrated as red spheres in **Figure 5.2**, suggest the three dimensional implications of primary structure truncations. The yCoq6 M469X truncation corresponds to the hCoq6 Q461fsX478 mutation. This mutant hCoq6 cannot complement Coq6-inactive strains of yeast.[5] However, the homologous mutation in yeast (M469), which presumably results in the deletion of the 11 C-terminal residues, retains C5-hydroxylation activity *in vivo*, but loses C4-deamination activity.[6]

The yCoq6 F455X mutation was mapped from the clinically documented hCoq6 W477X truncation mutation. The mutant hCoq6 is unable to complement Coq6-inactive strains of yeast[1] and neither can the equivalent yCoq6 F455X mutation.[4] This suggests that while the deletion of the 11 C-terminal yCoq6 residues is compatible with *in vivo* activity, truncations in the preceding (penultimate) alpha helix interfere with activity.

The yCoq6 W181X mutation was mapped from the clinically documented hCoq6 W188X truncation mutation, and the yCoq6 K155X mutation was mapped from hCoq6 R162X. Both of these mutations are much farther upstream in the protein's coding sequence, and yield inactive enzymes unable to complement yeast Coq6-inactive mutants.[1]

Aside from the truncation mutations, which likely provoke large changes to protein structure, a set of three point mutations (shown as cyan spheres in **Figure 5.2**) created only for the study of Coq6 seem to inactivate the enzyme by interfering with FAD binding.[7] These three mutations are G202V (which was implemented as a single point mutation) and G386A-N388D, which were implemented as a simultaneous double mutant. According to the yeast Coq6 RATIONAL model, yeast Coq6 G202 contacts the FAD pyrophosphate while forming a hairpin turn of highly conserved secondary structure with backbone angles unfavorable for any other residue. Mutation to valine likely alters the backbone conformation in this critical region involved in FAD binding, as well as providing a larger sidechain which may also sterically interfere with FAD binding. G386 forms part of the bottom of the FAD binding pocket, directly under the isoalloxazine with the alpha carbon pointing upwards towards it. The RATIONAL model suggests that addition of a methyl group at the G386 location could prevent the isoalloxazine from reaching a completely *in* conformation, thereby interfering with catalysis. N388, proximal to G386 also forms part of an alpha helix at the bottom of the FAD binding pocket, with its sidechain pointing downwards, away from the FAD. Our 3D model suggests that mutation to an aspartate could destabilize this helix, and also interfere with FAD binding.

The yCoq6 D201H mutation was mapped from the clinically documented hCoq6 D208H mutation.[8] This mutant hCoq6 cannot complement yeast Coq6-inactivated mutants, and its single-allele appearance in clinical cases results in a neuropathology. Mapped onto the RATIONAL Coq6 model, yCoq6 D201 also contacts the FAD directly at the adenine ring, at a position immediately adjacent to the G202 position described earlier.

The yCoq6 A361D mutation (mapped from hCoq6 A353D[1]) and the yCoq6 F420C mutation (mapped from hCoq6 Y412C[3]) are at surface exposed positions and are distal to the FAD binding and substrate binding regions. Nonetheless, these mutations also result in an inactive enzyme in humans. We note that the clinical hCoq6 mutations are reported as single-allele mutations; presumably a complete loss of Coq6 activity and consequent loss of endogenous Q biosynthesis is incompatible with multicellular life.

The last mutation in the table, yCoq6 G248R (mapped from hCoq6 G255R[1]) is the most interesting for our modeling of Coq6 because of its proximity to a structural feature identified in the previous chapter: the putative substrate access channel identified as *re* face tunnel 1.

2.1 The *H. sapiens* clinical mutation Coq6 G255R corresponds to *S. cerevisiae* Coq6 G248R

The human G255R mutation is of particular interest because its homologous mutation in yeast, G248R, occurs at the entrance of the putative substrate access channel in the RATIONAL Coq6 model developed in Chapter 4. yCoq6 G248 is at a surface exposed solvent accessible position. This glycine normally forms

part of a hairpin turn between two strands in the large beta sheet, and is solvent exposed. This gives it the steric freedom to tolerate the drastic mutation to arginine without severely disrupting the overall fold of the protein. We recall that G248 is immediately adjacent to P249, which was identified as a bottleneck residue for this tunnel (see **Figure 4.6** in Chapter 4). This suggested that a mutation to arginine at this position might block the tunnel. The initial build coordinates of the G248R mutation orient the arginine's sidechain towards the channel lumen, visibly creating a blockage of the tunnel.

3. MD and substrate docking of the G248R mutant

We decided to apply the same protocol of molecular dynamics and substrate docking developed for the wild-type yeast Coq6 to the analysis of the Coq6 G248R mutant. The principal difference is the *re* face tunnel 1 bottleneck diameter, now measured between the R248 sidechain and the L382 sidechain as shown in **Figure 5.3.**

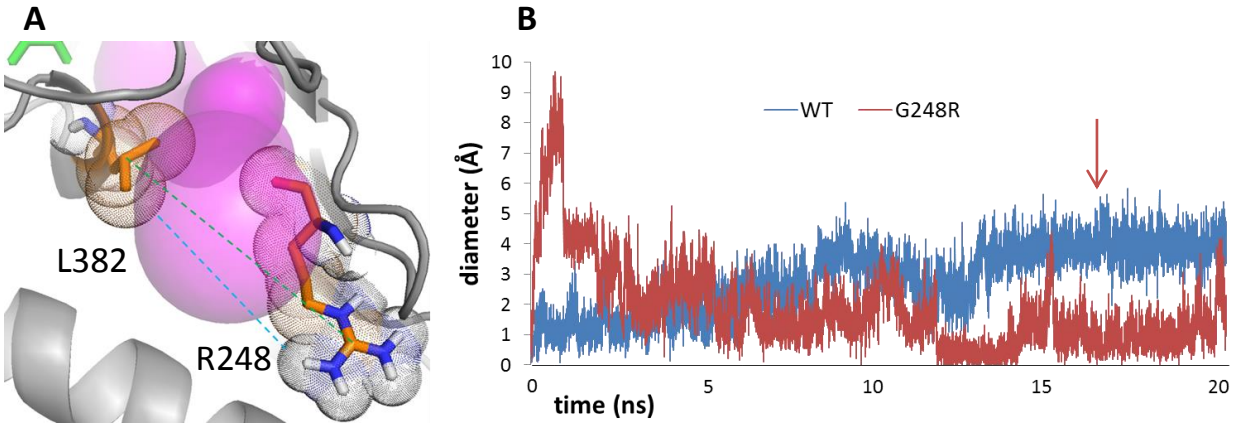

FIG 5.3: *A) Effective diameter of the re face tunnel 1 in the yeast Coq6 G248R mutant, as measured between the R248 CZ atom and the L382 CG atom, shown on the pre-MD RATIONAL model. The green line shows the center-to-center distance between the monitored bottleneck atoms. The blue line indicates the effective channel diameter corrected for the VDW radii. B) The effective diameter is usually too low to give a functional channel, but occasionally becomes large enough to admit passage of the substrate. The red arrow indicates the frame selected from G248R mutant model for substrate docking.*

As with the wild type models, this channel diameter metric enables us to select specific frames fro the molecular dynamics trajectories for substrate docking attempts. Arginine has a long and flexible sidechain which can adopt several conformations. For the purposes of admitting passage of the substrate to the active site, we can distinguish between two types of conformations of the arginine sidechain: blocking and non-blocking, illustrated iin **Figure 5.4** below. Panel A shows the geometry of the wild type P249 sidechain; Panels B and C show the G248R mutant in non-blocking and blocking conformations, respectively.

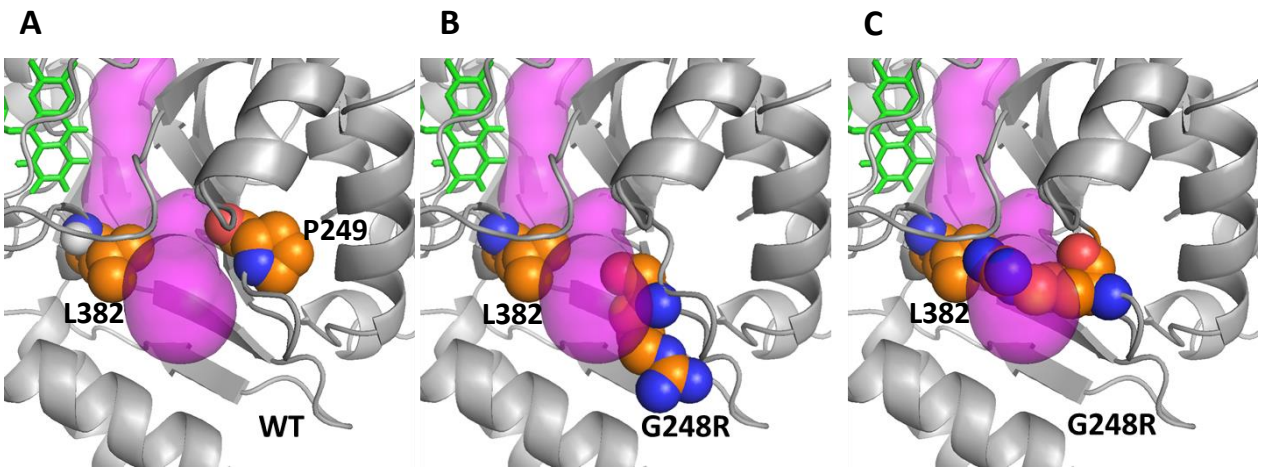

FIG 5.4: *Coq6 models illustrating the effect of the G248R mutant on re face tunnel 1. A) Coq6 wild-type (WT) model with the re face tunnel 1 (purple volume) bottleneck residues L382 and P249 shown as spheres. B) Coq6 G248R mutant model with the R248 sidechain in a non-blocking conformation. C) Coq6 G248R mutant model with the R248 sidechain in a tunnel blocking conformation.*

Guided by the active site geometry scoring function (Chapter 4, Section 8.2), docking of the model substrate (4-HP6, or 3-hexaprenyl-4-hydroxybenzoate) into non-blocking R248 conformations manages to recapitulate the substrate position found by docking into the wild type. An example of this result is shown in **Figure 5.5** below.

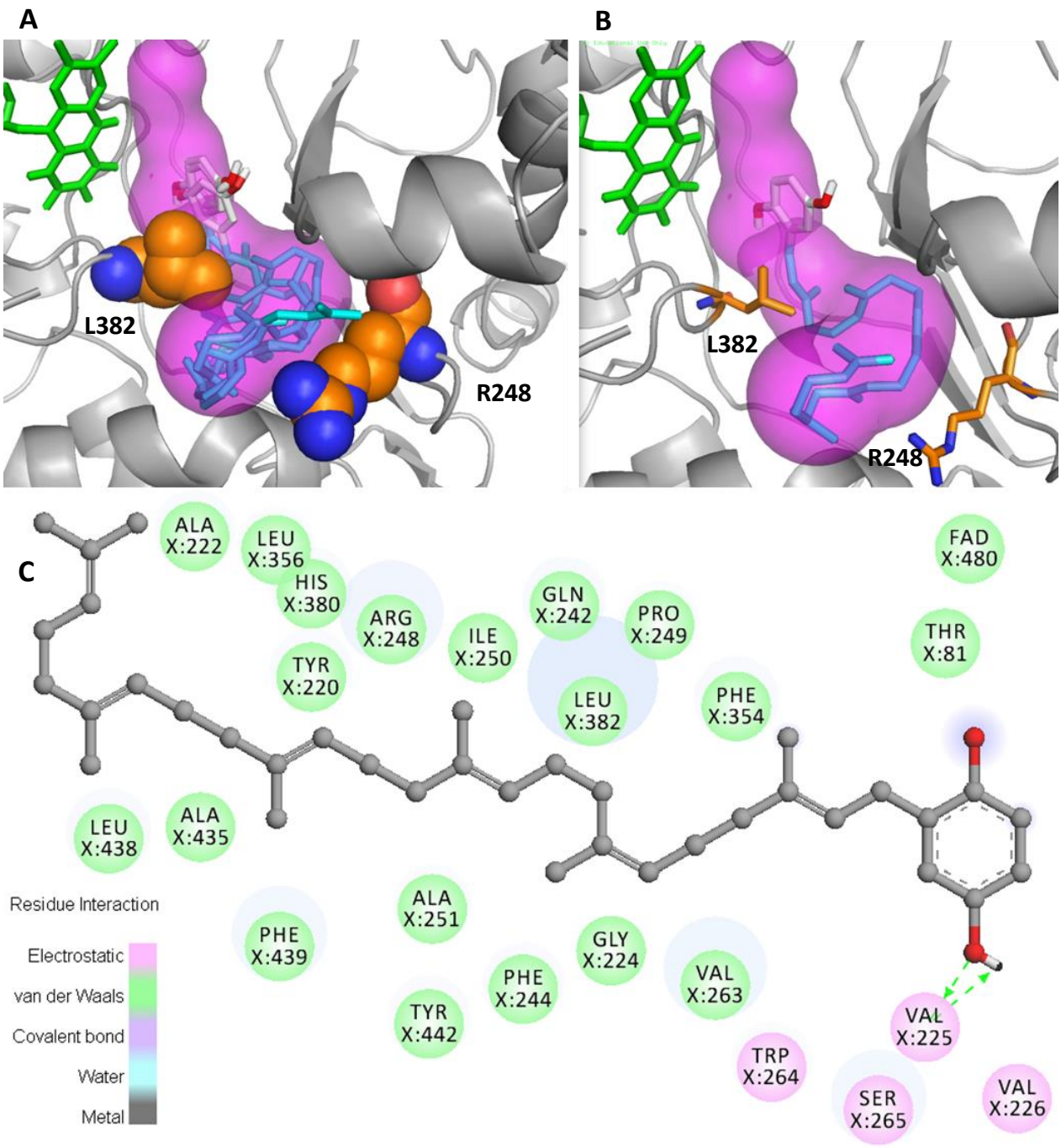

FIG 5.5: *Substrate docking into the Coq6 G248R model where the sidechain of R248 is in a non-blocking conformation, permitting passage of the polyprenyl tail of the substrate. This pose also places the aromatic head near the FAD isoalloxazine. A) The resulting poses from docking, showing a recurrence of the aromatic head positioning in front of the FAD isoalloxazine. B) A single pose selected with the smallest FAD C4a – substrate C5 distance (5.98 Å). C) An enzyme-substrate interaction plot for the selected docked substrate conformation.*

As we can see from **Figure 5.5C**, the polyprenyl tail forms many VDW contacts with hydrophobic and aromatic residues of re face tunnel 1, primarily small hydrophobics such as alanine and leucine which show strong evolutionary conservation (as shown in **Chapter 4, Figure 4.33**) While the aromatic head finds a position directly in front of the FAD isoalloxazine, its hydroxyl groups do find contacts with the same residues as in the RATIONAL model of the wild-type Coq6, as indicated by the tight clustering of the aromatic head poses in **Figure 5.5A**. Nonetheless, the distance between the FAD C4a atom (which will bear the reactive peroxo group) and the substrate's C5 atom is 4.78 Å, which is slightly larger than the 4.6 Å distance calculated in the WT model.

Overall, this is an encouraging result for our Coq6 model because it is consistent with the clinically observed phenotype of partially impaired Coq6 function resulting from the hCoq6 G255R mutation.[1] The human phenotype for this is steroid resistant nephrotic syndrome coupled with sensorineural deafness. This is interesting because these are pathologies in two cell types which may have exceptionally high requirements for endogenous Q biosynthesis. Both cell types are likely to require highly impermeable cell membranes for function (filtration in the renal podocytes affected in the nephrotic syndrome, and high ion pump activity in inner ear hair cells). Molecular dynamics shows us how this mutation can be tolerated by the Coq6 structure thanks to its surface exposed position, and how it can alternately block the substrate access channel or leave it open to the substrate. **This naturally occurring mutation gives us a good starting point for designing artificial mutations affecting the channel.**

4. Rational design of novel mutants blocking the substrate access channel

The ability of the G to R mutation at position 248 to partially block the substrate access channel while maintaining the global fold (as concluded from molecular dynamics and substrate docking) supports the structural validity of the RATIONAL Coq6 model in the substrate binding region and active site. It also strongly suggests that our prior identification of yeast Coq6 P249 (the position immediately adjacent) as a channel bottleneck is likely to be valid. Since the G248R mutation essentially gives us a test of the P249 position, we can now consider testing the other residue position implicated as a channel bottleneck: L382.

Several factors motivate us to carefully consider the mutation we will choose to test the important of position 382 as a channel bottleneck. First, we only have a limited laboratory capacity to generate and test any such mutants, which will motivate us to design a mutant with the highest probability of success on the first attempt. That is to say, we are seeking to produce a mutation which will maintain the global fold of the Coq6 enzyme as well as block the substrate access channel. Position 382 shares some similarities with position 248: it is at the entrance of the channel, the L382 sidechain points towards the channel lumen, yet it is also surface exposed. It is also diametrically opposed to the R248 sidechain, making interaction with it plausible. Finally, L382 is a highly conserved residue in the Coq6 family, and it forms part of a highly conserved structure of the global fold (one of the helices of the Rossmann fold, as shown in Chapter 3, **Figure 3.8**).

Knowing that our Coq6 homology model may diverge from the Coq6 experimental structure, we aim to propose a mutation that would also be robust to errors in our modeled coordinates. Preliminary tests using mutations to bulky hydrophobic residues (S267M, F439M) indicated that sidechain movement would not guarantee blockage of the tunnel. Indeed, the conformation of hydrophobic sidechains in the free volume of the channel lumen is mainly dependent on intrinsic rotamer preference. Instead, it would be useful to have some constraint on sidechain orientation to ensure channel blockage.

In the context of the naturally occurring hCoq6 G255R mutation (modeled in this work as yCoq6 G248R), we propose a geometric and electrostatic complementation by mutating L382 to glutamate. The glutamate sidechain has the same topology and roughly the same size as the wild-type leucine, indicating that it should be sterically accommodated by the position. Glutamate also has a carboxylate group on its sidechain, giving it a negative charge and therefore electrostatic complementarity to the R248 sidechain. This makes it possible to envisage the formation of a salt bridge between the two, as illustrated below in **Figure 5.6**.

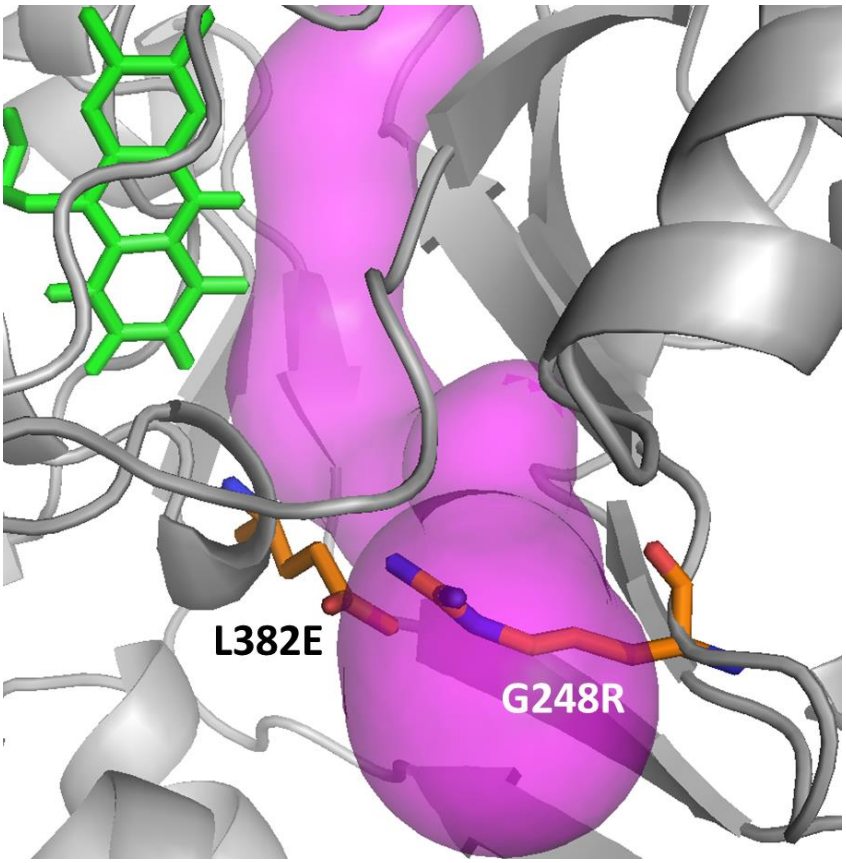

FIG 5.6: *Rational design of the G284R-L382E double mutation. The sidechains of positions L382 and R248 are oriented towards each other across the channel lumen. The electrostatic complementation is intended to maintain the R248 sidechain extended across the channel entrance to constitutively block it.*

4.1 MD and substrate docking of the L382E mutant

The next step in such a strategy is the modelling of the L382E mutation as performed for the G248R mutation: we will construct the homology model of the L382E single mutant, simulate its movements in molecular dynamics, select the best substrate binding conformations, and assess the traversability of the substrate channel through docking. The diameter of the L382E mutant's *re* face tunnel 1 is shown as the green trace in **Figure 5.7B**, compared to the diameter of the same tunnel in the wild-type (blue trace).

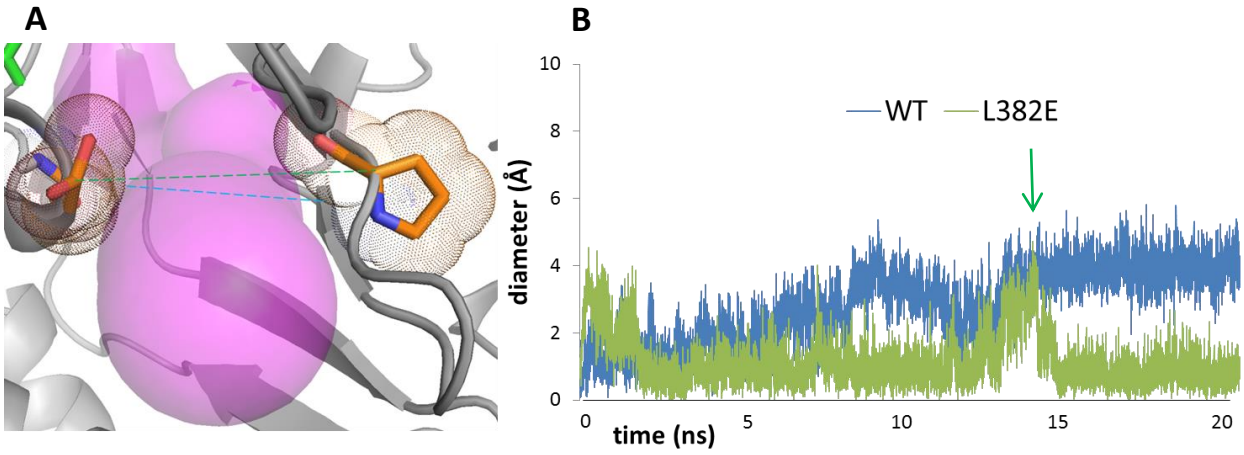

FIG 5.7: *A) Effective diameter of the re face tunnel 1 in the yeast Coq6 L382E mutant, as measured between the E382 sidechain and the P249 sidechain, shown on the pre-MD RATIONAL model. The green line shows the center to center distance between the monitored bottleneck atoms; the blue line indicates the effective channel diameter corrected by their VDW radii. B) The diameter is usually too low give a traversable channel, but occasionally becomes large enough to admit passage of the substrate. The green arrow indicates frame 6874 selected for substrate docking.*

Using the active site descriptor scoring function (presented in Chapter 4) in conjunction with the tunnel diameter criterion, we select a Coq6 conformation from the trajectory (indicated by the green arrow in **Figure 5.7B**) and test the traversability of the substrate access channel by docking, whose results are shown below. We note that the definition of the re face tunnel 1 bottleneck has changed: it is defined as occurring between P249 and E382, as illustrated in **Figure 5.7A**.

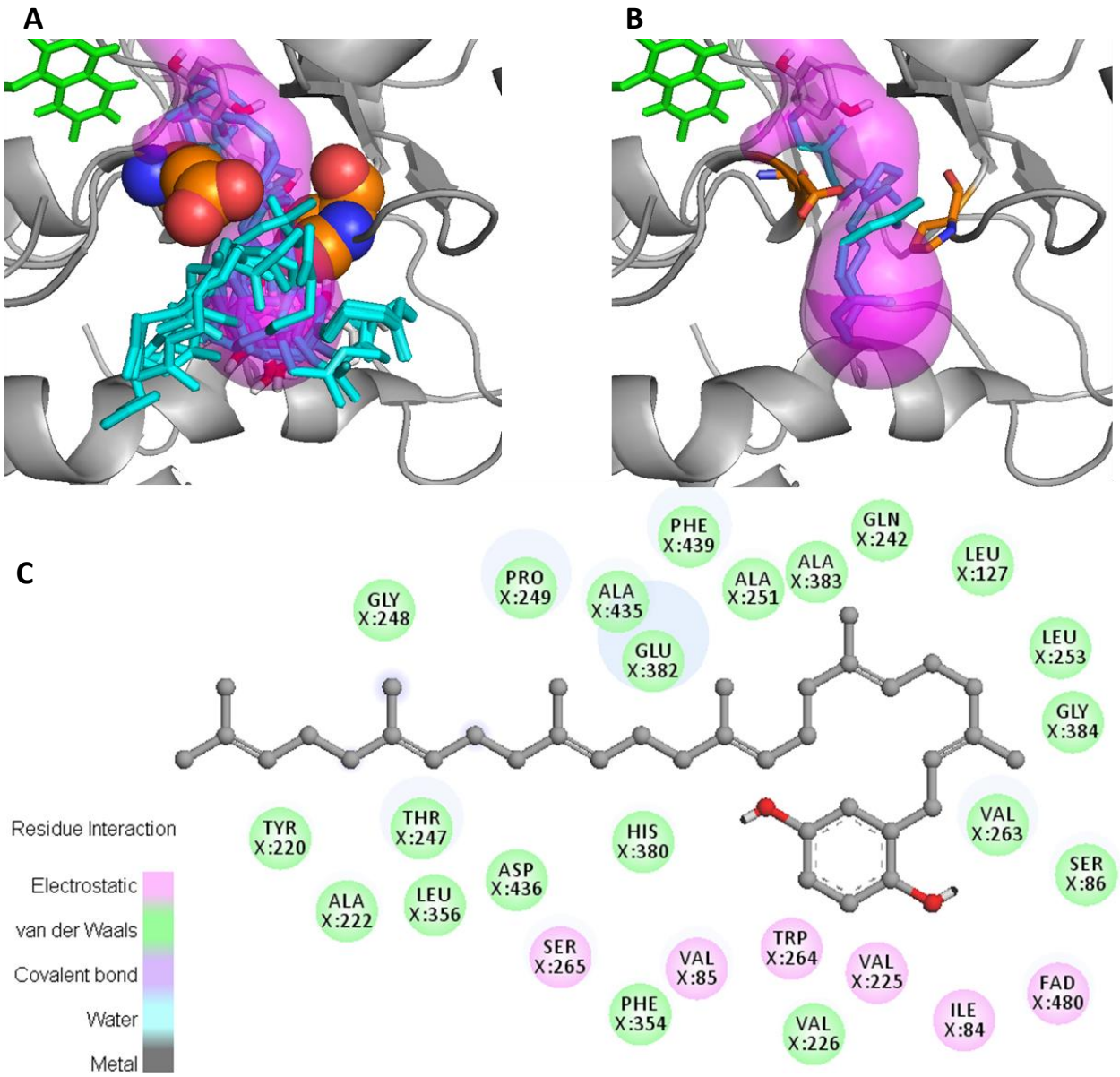

FIG 5.8: *Substrate docking into the RATIONAL Coq6 L382E model where the E382 sidechain is in a non-blocking conformation, permitting passage of the substrate's polyprenyl tail. This pose also places the aromatic head near the FAD isoalloxazine.*

The procedure manages to find enzyme conformations where the substrate can traverse the channel, including a docking pose where the substrate's aromatic head is placed in front of the FAD isoalloxazine. This suggests that the L382E mutation can be tolerated by the structure of the enzyme, and partially preserves the integrity of the substrate access channel. As can be seen from **Figure 5.7**, the L382E mutation greatly reduces the diameter of the channel, indicating it spends most of its time closed, switching to an open state only transiently.

4.2 MD and substrate docking of the G248R-L382E double mutant

The modeling results of both single point mutants G248R and L382E suggest that both mutations are positioned to have an effect on the substrate access channel by partially blocking it. Now we will attempt the complete and constitutive blocking of the channel by the formation of a salt bridge involving both the G248R and L382E mutations. Therefore, we will redefine the channel bottleneck as being between E382 and R248.

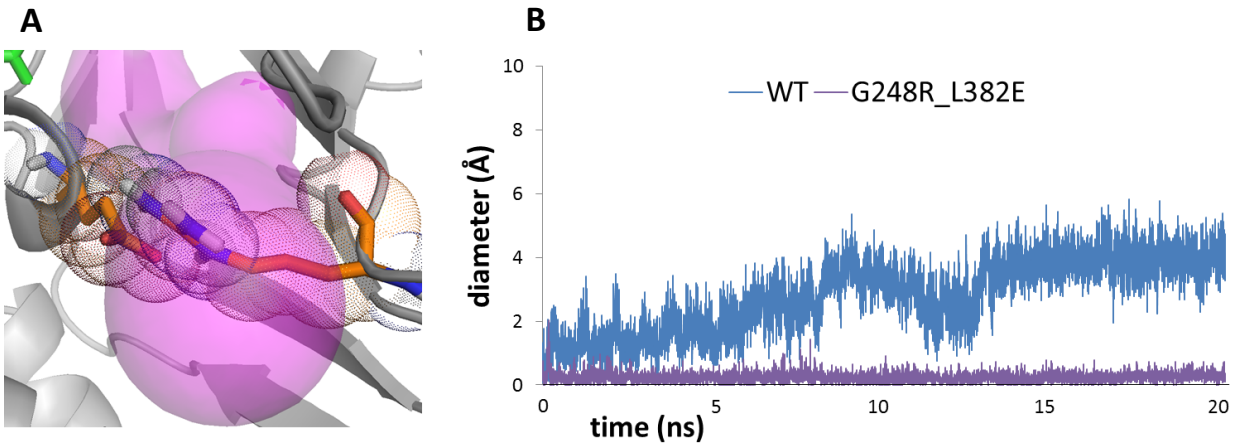

FIG 5.9: *A) Effective diameter of the re face tunnel 1 in the yeast Coq6 G248R-L382E double mutant, as measured between the E382 sidechain and the R248 sidechain, shown on the pre-MD RATIONAL model. The green line shows the center to center distance between the monitored bottleneck atoms; the blue line indicates the effective channel diameter corrected by their VDW radii. B) The salt bridge is stable over the course of the simulation, giving an effective channel diameter of nearly zero.*

The initial position of the E382 and R248 sidechains has them oriented towards each other across the channel entrance, allowing them to come into direct contact with each other. They maintain this contact through electrostatic attraction between their charged sidechains. This arrangement forms a stable salt bridge which persists over the 20ns of molecular dynamics, as illustrated in **Figure 5.9**. The effective diameter of the channel opening, as measured between the sidechains of these two residues, remains too small to allow passage of the substrate. Attempts at substrate docking into the active site failed, indicating that channel blockage caused by formation of the E382-R248 salt bridge is total. These results are resumed in **Figure 5.10** below, showing the effective diameters of re face tunnel 1 in the wild-type Coq6 model as compared to the rationally designed mutants.

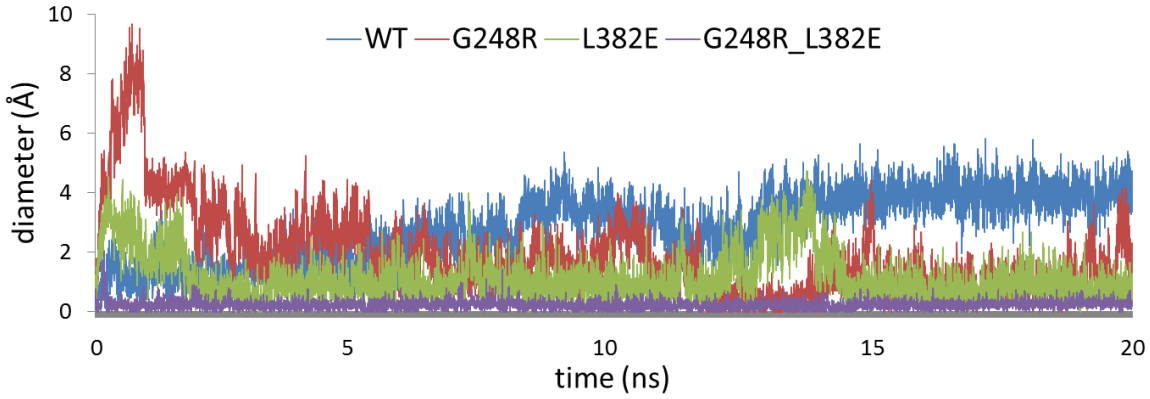

FIG 5.10: *Effective diameter of the re face tunnel 1 in the yeast Coq6 wild type and mutants. The wild type (blue trace) shows an average effective diameter of about 4Å during the latter half of the simulation. The single mutants (red and green traces) show much larger effective diameters which attain traversable values only occasionally. The double mutant, which forms a stable salt bridge across the tunnel, shows a diameter that is effectively near zero, constitutively blocking passage of the substrate.*

5. Experimental results

Thus far we have developed the general hypothesis of a substrate access channel in Coq6 and designed specific mutations to specific residues in order to block the passage of a model substrate (4-HP6, or 3-hexaprenyl-4-hydroxyphenol). The next step in our strategy is to test these mutations using *in vivo* activity assays. These were developed by our partner Dr. Fabien Pierrel (Université de Grenoble) and are presented here as a supporting evidence for the substrate access channel hypothesis.

5.1 *In vivo* activity assays for Coq6 WT, G248R, L382E, and G248R-L382E

The general principle of the assay relies on the functional complementation of *S. cerevisiae* Coq6-null mutants on two types of media: fermentable (glucose) and non-fermentable (lactate-glycerol). Non-fermentable media forces the yeast cells to use aerobic respiration for survival and growth, making them reliant on the ability to synthesize ubiquinone, which they can only produce if they have been complemented with a functional version of Coq6. In this assay, the exogenous Coq6 is supplied by transformation.

A key question about the designed mutations is their effect on the Coq6 protein structure. It is important to distinguish between a loss of Coq6 function due to specific blockage of the substrate access channel and loss of function due to general misfolding of the protein. The design goal of the channel blocking mutations was the former, and interestingly enough, the obligate nature of the CoQ synthome gives us a sensitive functional assay of the stability of Coq6 mutants. If the mutations have not disturbed the overall structure of the protein, it should still be able to participate in assembly of the CoQ synthome, producing a phenotype specifically deficient in C5 hydroxylation in Q biosynthesis. However, the correct assembly of the synthome should preserve the function of the other Q biosynthesis proteins, which can enable the rescue of C5 hydroxylation deficient mutants. This rescue of these deficient phenotypes is the second step of the assay, which will enable us to assess the proper folding of the Coq6 rationally designed mutants. These experimental results of testing the designed mutations are presented in the **Figure 5.11** below.

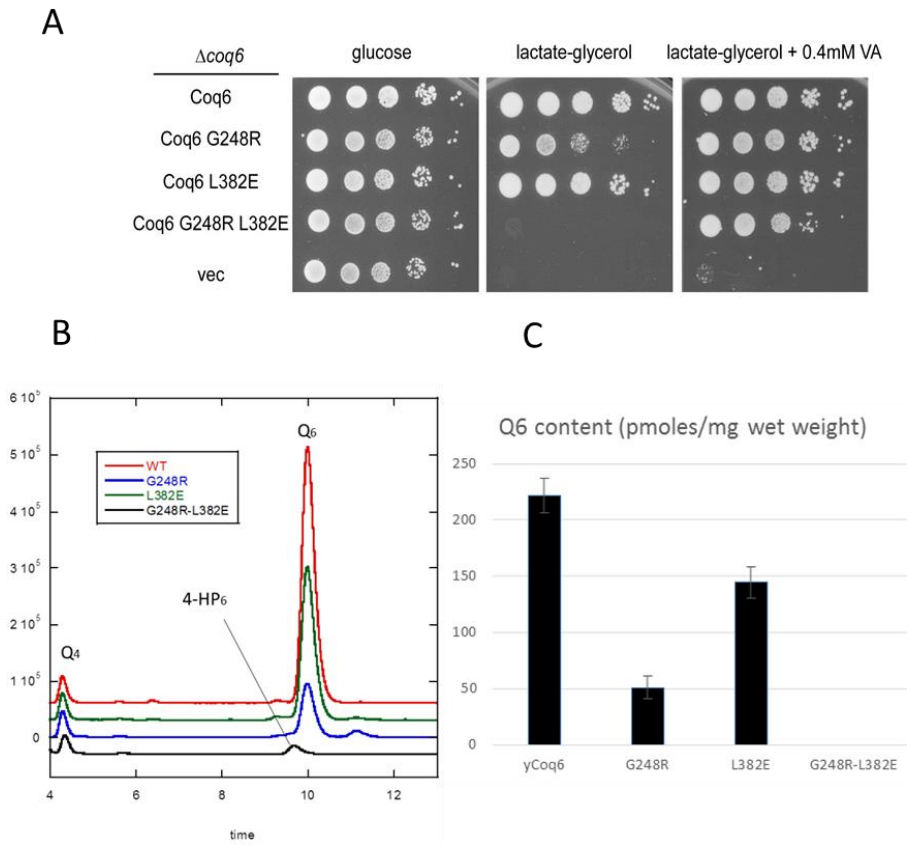

FIG 5.11: *A) 10 fold serial dilution of the Δcoq6 strain carrying an empty plasmid (vec) or plasmids coding for yCoq6, yCoq6-G248R, yCoq6-L382E and yCoq6-G248R-L382E. The plates contained YNB-pABA agar medium supplemented with the indicated carbon source and vanillic acid (VA) or not. The plates were imaged after incubation at 30°C for 2 days (glucose) or 6 days (lactate-glycerol). B) Representative electrochromatogram of lipid extracts from Δcoq6 cells expressing either yCoq6, yCoq6-G248R, yCoq6-L382E and yCoq6-G248R-L382E (1 mg of cells). The elution position of the Q_4 standard, of 3-hexaprenyl-4-hydroxyphenol (4-HP_6) and Q_6 are indicated. C) Q_6 amounts (in pmoles per mg of wet weight) in Δcoq6 cells expressing either yCoq6, yCoq6-G248R, yCoq6-L382E or yCoq6-G248R-L382E. Cells were grown in YNB–pABA (para-aminobenzoic acid, or 4-aminophenol) 2% lactate-glycerol containing 10 µM 4HB. The results are the average of 3-4 independent experiments and the error bars represent standard deviation.*

Section A of **Figure 5.11** is composed of three panels, each representing a different growth medium. The first panel shows a glucose medium supports aerobic metabolism as well as fermentation. The second panel shows a lactate-glycerol medium which forces the cells to use aerobic metabolism and cannot support fermentation. The third panel shows the same medium with the addition of vanillic acid, an unprenylated aromatic ring bearing a methoxy group on its C5 carbon and therefore a Q biosynthesis intermediate downstream of Coq6. If the CoQ synthome has been properly assembled, including a properly structured Coq6 enzyme, then addition of vanillic acid should enable rescue of the deficient phenotypes.

All mutants and the wild type are able to grow on the fermentable medium, because their growth is not strictly dependent on the endogenous biosynthesis of Q. However, significant differences are apparent for the mutants grown on the non-fermentable media presented in the second and third panels. The

quantities of Q synthesized by each mutant are measured by electrochemical detection, as presented in **Figure 5.11 panels B and C**.

Yeast colonies complemented by the various mutants of Coq6 are shown in each line on each plate. The first line shows the wild-type Coq6 with a fully functional Q biosynthesis system, which is able to grow on both fermentable and non-fermentable media. The second line shows the Coq6 G248R mutation, which shows reduced growth on non-fermentable media in the second panel, and rescued growth with the addition of vanillic acid in the third panel. These results indicate that the Coq6 G248R mutation reduces the activity of Coq6, but still allows the assembly of the CoQ synthome.

6. Conclusion

It is important to distinguish between a loss of Coq6 function due to specific blockage of the substrate access channel and loss of function due to general misfolding of the protein. The design goal of the channel blocking mutations was the former, and interestingly enough, the obligate nature of the CoQ synthome gives us a sensitive functional assay of the stability of Coq6 mutants. If the mutations have not disturbed the overall structure of the protein, it should still be able to participate in assembly of the CoQ synthome, producing a phenotype specifically deficient in C5 hydroxylation in Q biosynthesis, which was observed in the *in vivo* assay results. However, the correct assembly of the synthome should preserve the function of the other Q biosynthesis proteins, which can enable the rescue of C5 hydroxylation deficient mutants. This rescue of these deficient phenotypes is the second step of the assay, which implies that the rationally designed Coq6 mutants have structures near enough to the wild type to participate in the CoQ synthome.

Altogether, these *in vivo* results show that the G248R and L382E mutations decrease Coq6 activity to some extent while the combination of both mutations completely inactivates the enzyme without affecting its stability. These data are consistent with the theoretical prediction of the substrate channel being blocked by the proposed interaction between R248 and E382.

References

[1] Heeringa, Saskia F., Gil Chernin, Moumita Chaki, Weibin Zhou, Alexis J. Sloan, Ziming Ji, Letian X. Xie, et al. "COQ6 Mutations in Human Patients Produce Nephrotic Syndrome with Sensorineural Deafness." *The Journal of Clinical Investigation* 121, no. 5 (May 2011): 2013–24. doi:10.1172/JCI45693.

[2] Zhang, Keqiang, Jia-Wei Lin, Jinhui Wang, Xiwei Wu, Hanlin Gao, Yi-Chen Hsieh, Peter Hwu, et al. "A Germline Missense Mutation in COQ6 Is Associated with Susceptibility to Familial Schwannomatosis." *Genetics in Medicine* 16, no. 10 (October 2014): 787–92. doi:10.1038/gim.2014.39.

[3] Doimo, Mara, Eva Trevisson, Rannar Airik, Marc Bergdoll, Carlos Santos-Ocaña, Friedhelm Hildebrandt, Placido Navas, Fabien Pierrel, and Leonardo Salviati. "Effect of Vanillic Acid on COQ6 Mutants Identified in Patients with Coenzyme Q10 Deficiency." *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842, no. 1 (January 2014): 1–6. doi:10.1016/j.bbadis.2013.10.007.

[4] Ozeir, Mohammad, Ulrich Mühlenhoff, Holger Webert, Roland Lill, Marc Fontecave, and Fabien Pierrel. "Coenzyme Q Biosynthesis: Coq6 Is Required for the C5-Hydroxylation Reaction and Substrate Analogs Rescue Coq6 Deficiency." *Chemistry & Biology* 18, no. 9 (September 23, 2011): 1134–42. doi:10.1016/j.chembiol.2011.07.008.

[5] Heeringa, Saskia F., Gil Chernin, Moumita Chaki, Weibin Zhou, Alexis J. Sloan, Ziming Ji, Letian X. Xie, et al. "COQ6 Mutations in Human Patients Produce Nephrotic Syndrome with Sensorineural Deafness." *The Journal of Clinical Investigation* 121, no. 5 (May 2011): 2013–24. doi:10.1172/JCI45693.

[6] Ozeir, Mohammad, Ludovic Pelosi, Alexandre Ismail, Caroline Mellot-Draznieks, Marc Fontecave, and Fabien Pierrel. "Coq6 Is Responsible for the C4-Deamination Reaction in Coenzyme Q Biosynthesis in Saccharomyces Cerevisiae." *Journal of Biological Chemistry*, August 10, 2015, jbc.M115.675744. doi:10.1074/jbc.M115.675744.

[7] Ozeir, Mohammad, Ulrich Mühlenhoff, Holger Webert, Roland Lill, Marc Fontecave, and Fabien Pierrel. "Coenzyme Q Biosynthesis: Coq6 Is Required for the C5-Hydroxylation Reaction and Substrate Analogs Rescue Coq6 Deficiency." *Chemistry & Biology* 18, no. 9 (September 23, 2011): 1134–42. doi:10.1016/j.chembiol.2011.07.008.

[8] Zhang, Keqiang, Jia-Wei Lin, Jinhui Wang, Xiwei Wu, Hanlin Gao, Yi-Chen Hsieh, Peter Hwu, et al. "A Germline Missense Mutation in COQ6 Is Associated with Susceptibility to Familial Schwannomatosis." *Genetics in Medicine* 16, no. 10 (October 2014): 787–92. doi:10.1038/gim.2014.39.

Chapter 6 Research perspectives

1. Conclusion of the current work

The goal of molecular modeling of Coq6 has been to establish a structure-function characterization of the enzyme-cofactor-substrate system. The structural characterization has sought to answer three questions:

- What is the atomic resolution structure of Coq6?
- How does Coq6 bind its cofactor?
- How does Coq6 bind its substrate?

We have answered these three questions through the creation of a homology model of Coq6 and the characterization of the enzyme structure through a strategy of molecular dynamics, evolutionary residue conservation calculation, accessible volume detection, and docking of a rationally selected substrate model.

Homology modeling of the Coq6 sequence was non-trivial due to the low sequence identity between Coq6 and any of its templates, and further complicated by the conformational variety among the templates, and missing coordinates for major regions of some templates. These regions are functionally important, such as the FAD binding GDAxH loop (likely to be mal-formed in PDB structures 4K22 and 4N9X), and the substrate binding C-terminus (coordinates absent in 4K22 and 2X3N PDB structures). We addressed this through the creation of a panel of several homology models exploring several template combinations (using both a knowledge-based approach as well as the automated homology modeling servers I-TASSER and ROBETTA). We then tested their physical plausibility through molecular dynamics, leaving us with three stable models: one each from I-TASSER and ROBETTA, as well as a knowledge-based construction.

A realistic enzyme model should recapitulate molecular structure features important for enzymatic function. In the case of Coq6 it is easy to recognize the binding site for the enzyme's essential cofactor (FAD) as it is a fundamental feature of the enzyme's global fold. However, the binding site for the substrate is harder to recognize, and there are no experimentally solved structures of Q biosynthesis hydroxylases bound to substrates. Despite this, we have one key piece of information regarding the region of the protein likely to bind the substrate: the location of the FAD isoalloxazine moiety, which bears the reactive peroxo group necessary for catalysis. This is a position buried near the center of the protein structure, about 14Å away from the protein's exterior surface. In a well characterized enzyme of this general family (Class A flavoproteins), para-hydroxybenzoate hydroxylase (PHBH), it is known that a small aromatic substrate enters the buried active site through a channel which closes after substrate binding. This structural feature isolates the active site from the solvent, which is essential for catalysis in this family of enzymes. A functional Coq6 enzyme should also have a path for substrate access to the active site. However, the Coq6 substrates proposed in the literature are all hexaprenylated benzoquinones, with the covalently linked hexaprenyl tail accounting for most of the extended length of the chain (38.4Å) – a length longer than the burial depth (14 Å) of the active site. This strongly suggests that there must be a passage which allows the substrate, including at least part of its polyprenyl chain, to traverse from the exterior of the protein to the buried active site.

Our analysis of a partial crystal structure of a Q biosynthesis hydroxylase (PDB entry 4N9X) from another microorganism (*Erwinia carotovora*) solved in 2013 revealed the presence of a large channel connecting the surface to the active site (as computed with CAVER) composed of evolutionarily conserved residues (as computed with ConSurf). This strongly suggested that the functionally and structurally homologous Coq6 enzyme should have a similar feature, leading us to propose the hypothesis of a substrate access channel in Coq6.

We then analyzed each of the three remaining Coq6 models for accessible volumes and evolutionary conservation. These calculations, initially performed on the models before molecular dynamics, revealed three types of tunnels (appearing in all three models) which converge to the active site, located in front of the FAD. In light of our hypothesis of the existence of a substrate access tunnel, we decided to analyze the behavior of the tunnels over the molecular dynamics trajectories computed for each model. In this work we begin with the simplest hypothesis of enzyme-substrate binding which requires the fewest underlying hypotheses and initial conditions: conformational selection, as opposed to induced fit.

According to the hypothesis of a conformational selection mechanism, substrate-binding conformations of Coq6 should be accessible through molecular dynamics simulations without the substrate. We note that in any case the exploration of the conformational selection hypothesis must be performed before induced fit simulations. This is because the latter type of exploration requires initial positions for the substrate, which must be computed by considering the simpler case of conformational selection first.

Each tunnel was functionally characterized by its diameter at its narrowest point (as identified through accessible volume calculations) over the course of each trajectory. We selected conformations from each model's trajectory corresponding to maxima in tunnel bottleneck diameters, and then attempted docking a substrate model into these tunnels in order to assess their functional traversability.

Docking of a rationally selected substrate model (with our model substrate 3-hexaprenyl-4-hydroxybenzoate, also known as 4-HP6) revealed that only one model (the knowledge-based model) had a tunnel that could allow substrate access to the active site. This led us to select this model for further studies.

The proposed substrate access tunnel of this model was implicitly validated by the creation of a mutant Coq6 model (G248R) reproducing a clinically documented mutation in human Coq6 (G255R). This mutation is predicted to occur at the entrance of the putative substrate access channel and cause partial blockage of the channel. We rationally designed another mutation to the entrance of the substrate access channel (L382E). Alone, this mutation is also predicted to partially block the channel, which is supported by the reduced quantity of Q6 observed in strains bearing these mutant Coq6 constructs. When combined, these two mutations form a salt bridge across the tunnel entrance, constitutively blocking it and preventing substrate access. This is supported by the complete inability of yeast to survive on non-fermentable media when complemented with the Coq6 G248R-L382E double mutant. Together, modeling and experiment have been able to develop hypotheses and evidence for a substrate access channel in the Coq6 enzyme.

This work provides the first detailed structural information of an important and highly conserved enzyme of the Q biosynthesis pathway in the absence of crystallographic data. Our analysis indicates that in order to accommodate a bulky hydrophobic substrate Coq6 has evolved a substrate access channel to bind it and bring the aromatic head to a catalytic position in the active site. The availability of a structural model for Coq6 makes it possible to consider further computational approaches as detailed below.

2. Research perspectives

This project has been a good example of the value added of molecular modeling in the study of proteins that are difficult to purify, crystallize, and enzymatically characterize. Thus far, no *in vitro* activity assay exists for Coq6 for several reasons: low enzyme solubility, low substrate solubility, a complex redox system,[1] and possible inter-dependence on other enzymes of the CoQ biosynthesis complex.[2]

Molecular modeling can provide an alternate source of atomic coordinates for these complexes, enabling the formulation of structure-function hypotheses before the experimental resolution of structure. These coordinates of enzyme structure can enable the rational design of site-directed mutations, as well as provide a basis for a molecular understanding of experimentally observed phenotypes.

2.1 Molecular dynamics with substrate

Enzyme-substrate interactions have been the primary focus of the modeling performed in this work. The next step is to assess the stability of enzyme-substrate interactions found through ensemble docking by molecular dynamics simulations including the ligand.

2.2 Protein-protein interactions: binary pairs and protein complex architectures

Homology modeling of individual Coq proteins can be used to accelerate understanding of protein-protein interactions arising in this system. This includes *in vivo* protein interactions, important for understanding biological function, but also *in vitro* interactions, important for optimizing the purification process.

Our laboratory's work in overexpression and purification of Coq6 reveals that the protein is prone to aggregation during purification. This behavior was alleviated by adding a maltose-binding-protein tag for affinity purification, but returned when the tag was enzymatically cleaved. This implies that the aggregation tendency is an intrinsic property of the Coq6 enzyme. Preliminary calculations of the electrostatic surface of Coq6 indicate a bipolar charge distribution, giving the protein distinct negatively and positively charged faces. This could result in interactions which are long-range and powerful, but not specific in orientation or stoichiometry, leading to aggregation. Understanding the molecular structural basis of aggregation could help in designing new purification constructs or protocols.

Having atomic resolution molecular models is also useful for exploring protein-protein interactions likely to occur in the CoQ synthome. Developing a homology model for a potential partner protein (as identified through experimental interaction assays, for example) can enable protein-protein docking to explore protein-protein interactions in the complex. This approach is likely to require at least some experimentally derived distance restraints, as the results of protein-protein docking are quite sensitive to the conformation of the protein, both of the global fold, but also particularly of the surface residues.

Possessing homology models of each Coq protein can also be useful in the reconstruction of the CoQ complex or any relevant subcomplexes.[3] The approach for this combines two types of structural data: large scale, low resolution electron density maps of crystallized CoQ synthome particles (typically obtained through cryogenic electron-microscopy), and smaller scale, atomic resolution structures. The large scale (but low resolution) map of the protein complex is used as a constraining volume for fitting the high-resolution structures of its experimentally determined constituent proteins, yielding a high resolution structure of a macromolecular object normally too large to resolve as a single entity.

2.3 Protein-membrane interactions

Molecular simulation also makes it possible to explore interactions between proteins and membranes. While the mature Coq6 enzyme does not contain obvious membrane association domains, it may interact with the membrane in the context of the *in vivo* protein complex. Again, setting up this type of simulation is likely to require experimentally derived identification of residues likely to interact with the membrane, as there are many possible relative orientations between protein and membrane.

2.4 A phylogenetic study of the evolution of the Coq6-family insert

The application of molecular modeling to the Coq6 system has been largely focused on enzyme-substrate interactions, but there are other research avenues to explore. A major distinguishing feature of the Coq6 family of enzymes is the presence of the large insert region which has no experimentally solved structural homolog.

The insert region is of intrinsic structural interest because it is clearly not required for catalysis in the homologous bacterial enzymes we used as templates. Our work has suggested that the insert likely evolved from the elaboration of an existing element of secondary structure, an alpha helix on the outer surface of the beta sheet domain. A multiple sequence alignment of the Coq6 family enzymes shows that the insert is highly variable in composition and length. The ConSurf method, which was used to compute the MSA, also produces a phylogenetic tree which could be used in such a study.

We can envisage a more systematic study of the structure of the insert in the context of the Coq6 proteins by making and testing homology models for representative Coq6 sequences. Just as we did for *S. cerevisiae* Coq6, secondary structure prediction can be used to generate secondary structure assignments for the insert during the model building process, and molecular dynamics can be used to assess their structural stability. Performing this systematically on a selection of Coq6 family proteins can help us learn about the evolution of this structure.

2.5 Modeling of C-terminal truncation mutants: a role in the deamination of 3-hexaprenyl-4-aminobenzoate?

The C-terminus of the Coq6 protein has been shown to be important for deamination of C4-aminated substrates, but not for the C5 hydroxylation of C4-hydroylated substrates. Substrate docking results into the full length wild type Coq6 from *S. cerevisiae* indicate that the polyprenyl tail can play a role in the shielding of the active site from bulk solvent, particularly in the structural context of the CoQ synthome. There is also experimental evidence for interaction with Coq9 through the Coq6 C-terminus.[4] This suggests that contact with Coq9 may help seal the active site as well. Nonetheless, it would be interesting to see how truncation of the 11 C-terminal residues affects the isolated protein structure.

2.6 Molecular dynamics over longer timescales

The molecular dynamics simulations used in this project were 20 ns long. This was long enough to observe structural stabilization of some models, and long enough to allow sufficient conformational sampling of the models' tunnel systems. In our particular case, the best homology model indicated that the putative substrate access tunnel's transition between open and closed is mainly a function of sidechain motion, not requiring larger scale motions of secondary or super-secondary structural elements. However, it would still be of interest to run longer simulations in order to investigate motions that may be implicated in function.

2.7 Substrate-enzyme assignment through systematic molecular modeling

An open question about the Q biosynthesis pathway is the order of the reactions, as well as the attribution of specific intermediates to specific enzymes as substrates or products. Although several biosynthesis pathways have been proposed in the literature, the substrate-enzyme assignments have not been made on the basis of *in vitro* testing of specific substrates in controlled conditions. Indeed, at least one enzyme, Coq2, is known to be able to prenylate Q biosynthesis intermediates at varying levels of aromatic ring substitution. If other enzymes can also display substrate promiscuity, there may not be a single order for the biosynthesis pathway.

The structural and functional interdependence of Q biosynthesis proteins makes it difficult to make this assignment based on the accumulation of intermediates in functional knockouts or knock-downs. Thus far, the main tool for stabilizing Coq null mutants and detecting some intermediates has been overexpression of Coq8. Molecular modeling offers another powerful tool that can be used is a systematic way to determine enzyme-substrate attribution through accumulation of intermediates. As we have seen with the modeling of Coq6, it is possible to identify the active site and the substrate binding site in terms of explicit residues. This knowledge makes it possible to catalytically inactivate Coq proteins without disrupting their tertiary structure and incorporation into the CoQ synthome.

The strategy of such an approach would be to make homology models of all Coq enzymes in order to identify their catalytic residues. For each enzyme, we can propose inactivating mutations, and create mutant proteins for functional assays in an *in vivo* system, similar to the one used by Dr. Pierrel in this work. Each inactivated mutant Coq enzyme can be used to complement its respective null mutant, and each resulting strain should then accumulate intermediates diagnostic of the blocked reaction. This is a powerful strategy and technique for determining enzyme-substrate attribution using existing tools, long before the experimental structural resolution of the Coq enzymes.

References

[1] Pierrel, Fabien, Olivier Hamelin, Thierry Douki, Sylvie Kieffer-Jaquinod, Ulrich Mühlenhoff, Mohammad Ozeir, Roland Lill, and Marc Fontecave. "Involvement of Mitochondrial Ferredoxin and Para-Aminobenzoic Acid in Yeast Coenzyme Q Biosynthesis." *Chemistry & Biology* 17, no. 5 (May 28, 2010): 449–59. doi:10.1016/j.chembiol.2010.03.014.

[2] Allan, Christopher M., Agape M. Awad, Jarrett S. Johnson, Dyna I. Shirasaki, Charles Wang, Crysten E. Blaby-Haas, Sabeeha S. Merchant, Joseph A. Loo, and Catherine F. Clarke. "Identification of Coq11, a New Coenzyme Q Biosynthetic Protein in the CoQ-Synthome in Saccharomyces Cerevisiae." *Journal of Biological Chemistry*, January 28, 2015, jbc.M114.633131. doi:10.1074/jbc.M114.633131.

[3] Yang, Zheng, Keren Lasker, Dina Schneidman-Duhovny, Ben Webb, Conrad C. Huang, Eric F. Pettersen, Thomas D. Goddard, Elaine C. Meng, Andrej Sali, and Thomas E. Ferrin. "UCSF Chimera, MODELLER, and IMP: An Integrated Modeling System." *Journal of Structural Biology*, Structural Bioinformatics, 179, no. 3 (September 2012): 269–78. doi:10.1016/j.jsb.2011.09.006.

[4] Ozeir, Mohammad, Ludovic Pelosi, Alexandre Ismail, Caroline Mellot-Draznieks, Marc Fontecave, and Fabien Pierrel. "Coq6 Is Responsible for the C4-Deamination Reaction in Coenzyme Q Biosynthesis in Saccharomyces Cerevisiae." *Journal of Biological Chemistry*, August 10, 2015, jbc.M115.675744. doi:10.1074/jbc.M115.675744.

Annex 1 Preliminary Coq protein modeling

The initial results of modeling queries submitted to the Phyre2 server for HMM based template searching. Because of space considerations, the results for each Coq protein are presented in a separate electronic file.

Annex 2 Coq6 family multiple sequence alignment by ConSurf

The multiple sequence alignment computed by ConSurf on the basis of the Coq6 amino acid sequence Because of space considerations, the results for each Coq protein are presented in a separate electronic file.

Annex 3 Template study: Accessible volumes of the 2X3N structure

1. Introduction

The 2X3N PDB entry is the structure of the PqsL enzyme from *Pseudomonas aerguinosa*. The PqsL enzyme is from the 4-quinolone biosynthesis pathway in this organism.[1] [2] Quinolones are used as signaling molecules in this organism and many other bacteria.[3] [4] PqsL has been shown to be required for the biosynthesis of HQNO, (2-heptyl-4-hydroxyquinoline-N-oxide) shown in the figure below.[5]

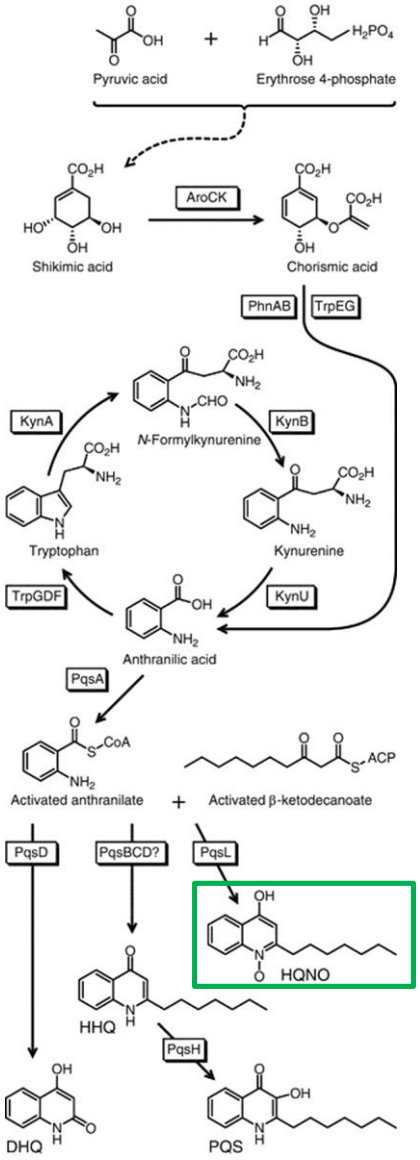

FIG A3.1: *Quinolone biosynthesis pathway proposed by Heeb et al (2011).*[2] *The likely product of PqsL is HQNO, shown in the green box.*

In addition to the structural similarity to ubiquinone (and even more structurally similar to menaquinones), HQNO has also been shown to be functionally similar, able to perform the roles of ubiquinone and menaquinone with cytochrome b enzymes from several organisms. [6 7 8]

Below, we show residue conservation and accessible volume calculations for the 2X3N structure. Residues from the FAD binding pocket are scored as highly conserved as shown by their purple coloring. Also shown are the tunnels computed with CAVER, which reveal three main types of tunnels which converge to the active site: *re* face tunnel 1 (purple volume), a possible *re* face tunnel 2 (blue volume), and a *si* face tunnel (red volume). However, as is visible in Panel D, none of these tunnels correspond to large regions conserved through evolution.

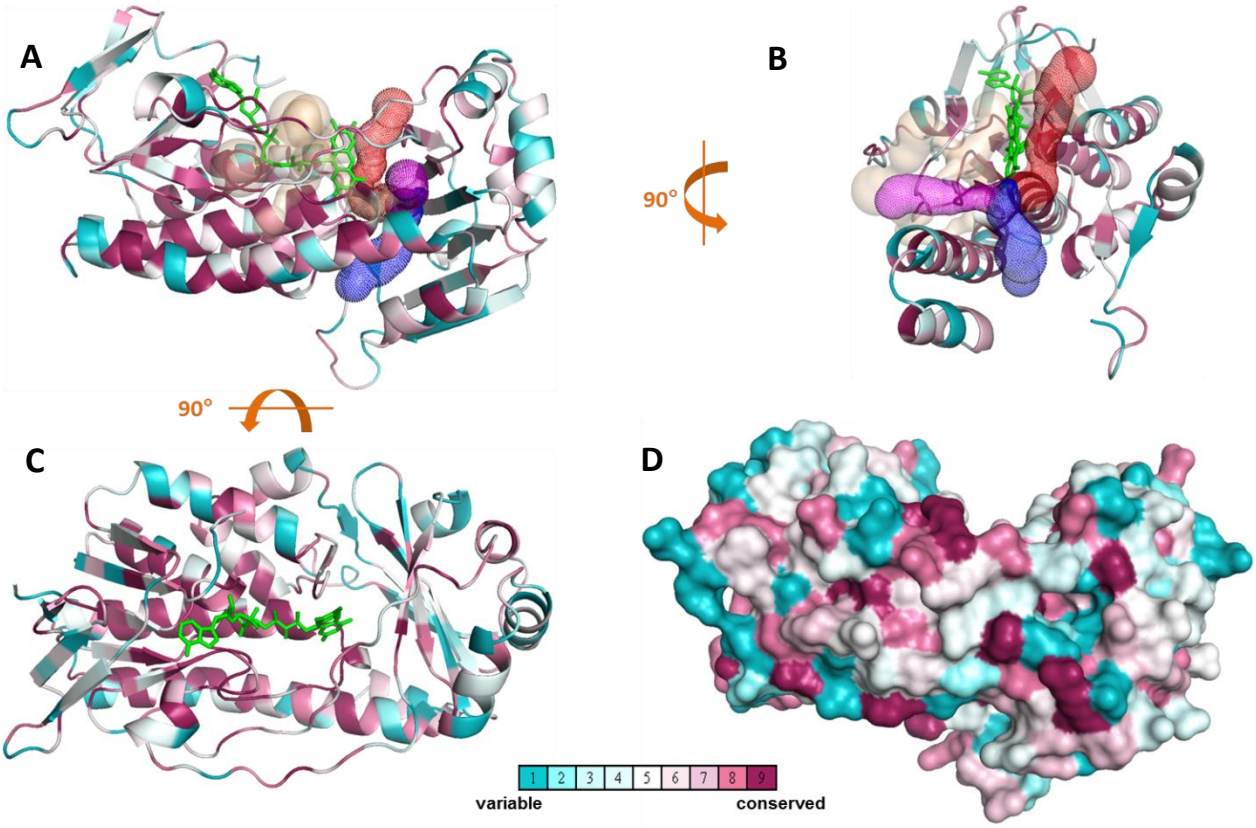

FIG A3.2: *Evolutionary residue conservation calculated by ConSurf from the 2X3N PDB structure of PqsL. The backbone cartoon is colored by the ConSurf color scheme. FAD is shown in green stick. Also shown is the tunnel system calculated by CAVER. Ghost tunnels are shown in beige. This structure also shows three main types of tunnels recurrent in enzymes of this global fold: re face tunnel 1 (purple), another tunnel which may exit via the re face in blue (but in 2X3N the C-terminus residues are not resolved), and a si face tunnel (red). A) view from the re face, B) view through the sheet face, C) view from the "top" of the FAD binding pocket, tunnels omitted for clarity, D) residue conservation projected onto the surface of the 2X3N structure.*

References

[1] Essar, D. W., L. Eberly, A. Hadero, and I. P. Crawford. "Identification and Characterization of Genes for a Second Anthranilate Synthase in Pseudomonas Aeruginosa: Interchangeability of the Two Anthranilate Synthases and Evolutionary Implications." *Journal of Bacteriology* 172, no. 2 (February 1990): 884–900

[2] Heeb, Stephan, Matthew P. Fletcher, Siri Ram Chhabra, Stephen P. Diggle, Paul Williams, and Miguel Cámara. "Quinolones: From Antibiotics to Autoinducers." *FEMS Microbiology Reviews* 35, no. 2 (March 2011): 247–74. doi:10.1111/j.1574-6976.2010.00247.x.

[3] Huse, Holly, and Marvin Whiteley. "4-Quinolones: Smart Phones of the Microbial World." *Chemical Reviews* 111, no. 1 (January 12, 2011): 152–59. doi:10.1021/cr100063u.

[4] Taylor, G. W., Z. A. Machan, S. Mehmet, P. J. Cole, and R. Wilson. "Rapid Identification of 4-Hydroxy-2-Alkylquinolines Produced by Pseudomonas Aeruginosa Using Gas Chromatography-Electron-Capture Mass Spectrometry." *Journal of Chromatography. B, Biomedical Applications* 664, no. 2 (February 17, 1995): 458–62.

[5] Hoffman, Lucas R., Eric Déziel, David A. D'Argenio, François Lépine, Julia Emerson, Sharon McNamara, Ronald L. Gibson, Bonnie W. Ramsey, and Samuel I. Miller. "Selection for Staphylococcus Aureus Small-Colony Variants due to Growth in the Presence of Pseudomonas Aeruginosa." *Proceedings of the National Academy of Sciences* 103, no. 52 (December 26, 2006): 19890–95. doi:10.1073/pnas.0606756104.

[6] Van Ark, Gerrit, and Jan A. Berden. "Binding of HQNO to Beef-Heart Sub-Mitochondrial Particles." *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 459, no. 1 (January 6, 1977): 119–37. doi:10.1016/0005-2728(77)90014-7.

[7] Smirnova, Irina A., Cecilia Hägerhäll, Alexandre A. Konstantinov, and Lars Hederstedt. "HOQNO Interaction with Cytochrome B in Succinate:menaquinone Oxidoreductase from Bacillus Subtilis." *FEBS Letters* 359, no. 1 (February 6, 1995): 23–26. doi:10.1016/0014-5793(94)01442-4.

[8] Richard A. Rothery, Joel H. Weiner. "Interaction of an Engineered [3Fe-4S] Cluster with a Menaquinol Binding Site of Escherichia Coli DMSO Reductase †." *Biochemistry* 35, no. 10 (1996): 3247–57. doi:10.1021/bi951584y.

Annex 4 Active site geometry descriptor test: redocking pHB into PHBH

1. Introduction

The development of a geometric descriptor of the active site used to characterize Coq6 was first developed and tested on PHBH structure 1PBE, which includes FAD and the substrate. The binding mode of the PHBH substrate (para-hydroxybenzoate hydroxylase) was analyzed in the 1PBE structure as a reference for the identification of active site conformations compatible with substrate binding. A set of enzyme atoms participating in hydrogen bonds with the substrate were identified, and distances between only the enzyme atoms were measured using VMD. This set of reference distances describes a 3D fingerprint of the enzyme active site when it has bound a substrate in a catalytically competent pose. We then tracked the interatomic distances between these atoms over the course of the MD trajectory. In order to quantify the geometric similarity between the active site conformations before MD and those sampled during MD we created a scoring function based on the differences between the interatomic distances measured before and during MD. This function is shown below.

$$S = \Sigma\ [d_{ij}(1PBE) - d_{ij}(1PBE)]$$

The PHBH residues binding the pHB substrate are as follows: the substrate's hydroxyl group is hydrogen bonded to three residues (Y201, P293, T294). The substrate's carboxyl group forms hydrogen bonds with different residues (S212 R214, Y222). These bonds are illustrated in **Figure 1** below.

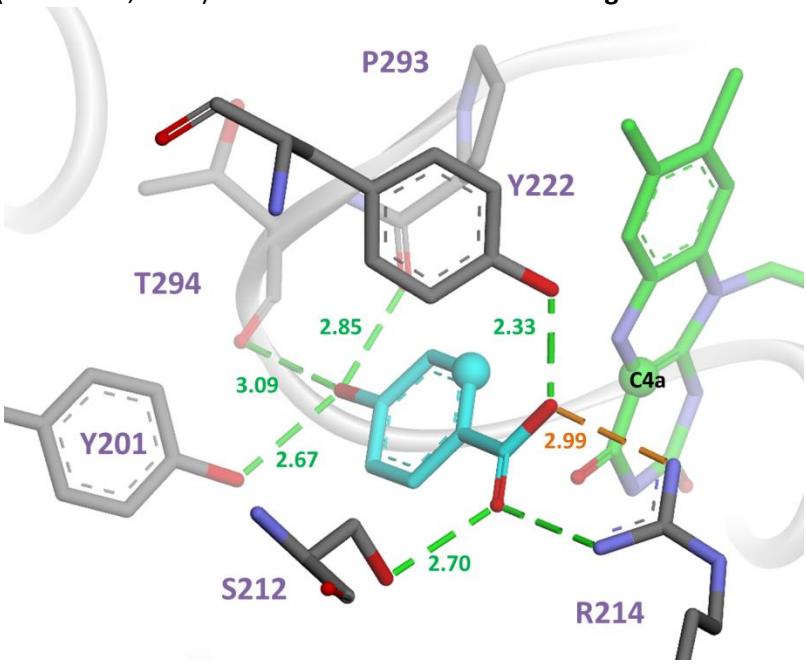

FIG A4.1: *The active site of PHBH structure 1PBE. The substrate, para-hydroxybenzoate, is shown in cyan stick with its hydroxylation target carbon shown as a cyan sphere. The FAD is shown in green, with its C4a carbon (which will bear the reactive peroxo group) shown as a green sphere. Residues binding the substrate through hydrogen bonds are shown in grey sticks. The substrate's hydroxyl group is hydrogen bonded to three residues (Y201, P293, T294), two of which (P293 and T294) bond through backbone*

oxygens. The substrate's carboxyl group forms hydrogen bonds with different residues (S212, R214, Y222). Most notable is R214, presenting the bidentate guanidinium group to bind the carboxyl oxygens. However, only a subset of these possible distances were measured. These particular distances are shown in **Figure 2** below.

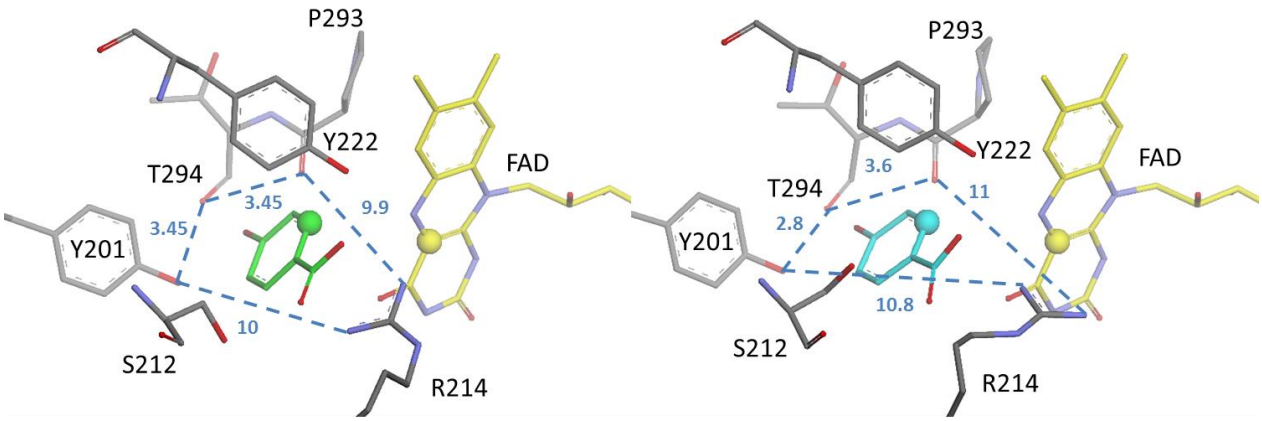

FIG A4.2: *A) The active site of PHBH structure 1PBE showing the receptor based interatomic distances used in creating the geometric descriptor and scoring function. B) The active site of a PHBH structure sampled from MD showing the distances between selected substrate binding atoms.*

This is because most of the computational work of molecular dynamics goes to simulating thermal motion. This means that the trajectories we analyze for signals of functional movements are largely dominated by the noise of thermal motion. The practical result is that when we are analyzing the relative motions of atoms in the simulation, we must be very picky about the atoms we choose to monitor. We must choose enough coordinates to capture the geometry of interest, but not more than that or we will introduce more and more thermal noise into our measurements, which will make trajectory analysis results more difficult to interpret.

Therefore we selected the subset of atoms indicated in **Figure A4.2** and use the scoring function to select frames from dynamics that most resemble the substrate-bound crystal structure.

Annex 5 Generation 2 Coq6 models

1. Homology models without the insert are used to test the stability of various constructs

Our review of the top templates has helped us define the templates we will use for the N-terminal and C-terminal regions of Coq6, while defining the precise sequence and position of the insert will tell us which Coq6 residue we can safely omit from this set of models. We have two choices for the N-terminus (4K22 and 2X3N) and two choices for the C-terminus (1PBE and 4N9X). Giving us four possible constructions. In addition, we also submitted the Coq6 sequence to the I-TASSER server to obtain independently generated models based on either the 4K22 or 2X3N structure. These possibilities are resumed in **Table A5.1 below**.

TABLE A5.1: *Table of the multi-template homology model coordinate sources for the N-terminal and C-terminal regions of Coq6. Here we present the coordinate sources for Generation 2 Coq6 models.*

Coq6 Target sequence	N-ter	C-ter
Generation 2 Without Insertion	2X3N	1PBE
	2X3N	4N9X

2. Construction of Generation 2 models without the insert

First we will present the alignments used to construct the Gen2 models of Coq6 without the insertion. Then we will compare the resulting homology models before comparing their behavior in molecular dynamics simulations.

```
         25         35         45         55         65         75
Coq6_WI PKLTDVLIVG GGPAGLTLAA SIKNSPQLKD LKTTLVDMVD LKDKLSDFYN SPPDYFTNRI
2X3N    -NHIDVLING CGIGGAMLAY LLGR--QGH- -RVVVVEQAR RE-------- -------NGA
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

         85         95        104        114        124        130
Coq6_WI VSVTPRSIHF LENNA-GATL MHDRIQSYDG LYVTDGCSKA TLDLA----R DSMLCMIEII
2X3N    DLLKPAGIRV VEAAGLLAEV TRRGGRVRHE LEVYHDGELL RYFNYSSVDA RGYFILMPCE
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

         140        150        160        170        180        190
Coq6_WI NIQASLYNRI SQYDSKKDSI DIIDNTKVVN IKHSDPNDPL SWPLVTLSNG EVYKTRLLVG
2X3N    SLRRLVLEKI DGEAT----V EMLFETRIEA VQRDE---RH AIDQVRLNDG RVLRPRVVVG
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

         200         210        219        229        239        249
Coq6_WI ADGFNSPTRR -FSQIPSRGW MYNAYGVVAS MKLEYPPFKL RGWQRFLPTG PIAHLPMPEN
2X3N    ADGIASYVRR RLLDIDVERR PYPSPMLVGT FALAPCVAER NRLYVDSQGG LAYFYPIGFD
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

         259        269        279        288          344      352
Coq6_WI NATLVWSSSE RLSRLLLSLP P-ESFTALIN AAFVLEDA-- --VSIIDKTR ARFPLKLTHA
2X3N    RARLVVSFPR EEARELMADT RGESLRRRLQ RFVGDESAEA IAAVTGTSRF KGIPIGYLNL
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

         362        372        382        392        402        412
Coq6_WI DRYCTDRVAL VGDAAHTTHP LAGQGLNMGQ TDVHGLVYAL EKAMERGLDI GSSLSLEPFW
2X3N    DRYWADNVAM LGDAIHNVHP ITGQGMNLAI EDASALADAL DLALRDACA- -LEDALAGYQ
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

         422        432        442        452        462        472
Coq6_WI AERYPSNNVL LGMADKLFKL YHTNFPPVVA LRTFGLNLTN KIGPVKNMII DTLGGNEK--
2X3N    AERFPVNQAI VSYGH----- ---------- ---------- ---------- ----------
1PBE    ---------- -----WMTSV LHRFPDTD-A FSQRIQQTEL EYYLGSEAGL ATIAENYVGLP
```

FIG A5.1 *Construction alignment of Gen2 WT WI 2X3N_1PBE*

```
         25         35         45         55         65         75
Coq6_WI PKLTDVLIVG GGPAGLTLAA SIKNSPQLKD LKTTLVDMVD LKDKLSDFYN SPPDYFTNRI
2X3N    -NHIDVLING CGIGGAMLAY LLGR--QGH- -RVVVVEQAR RE-------- -------NGA
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

         85         95        104        114        124        130
Coq6_WI VSVTPRSIHF LENNA-GATL MHDRIQSYDG LYVTDGCSKA TLDLA----R DSMLCMIEII
2X3N    DLLKPAGIRV VEAAGLLAEV TRRGGRVRHE LEVYHDGELL RYFNYSSVDA RGYFILMPCE
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

         140        150        160        170        180        190
Coq6_WI NIQASLYNRI SQYDSKKDSI DIIDNTKVVN IKHSDPNDPL SWPLVTLSNG EVYKTRLLVG
2X3N    SLRRLVLEKI DGEAT----V EMLFETRIEA VQRDE---RH AIDQVRLNDG RVLRPRVVVG
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

         200         210        219        229        239        249
Coq6_WI ADGFNSPTRR -FSQIPSRGW MYNAYGVVAS MKLEYPPFKL RGWQRFLPTG PIAHLPMPEN
2X3N    ADGIASYVRR RLLDIDVERR PYPSPMLVGT FALAPCVAER NRLYVDSQGG LAYFYPIGFD
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

         259        269        279        288          344      352
Coq6_WI NATLVWSSSE RLSRLLLSLP P-ESFTALIN AAFVLEDA-- --VSIIDKTR ARFPLKLTHA
2X3N    RARLVVSFPR EEARELMADT RGESLRRRLQ RFVGDESAEA IAAVTGTSRF KGIPIGYLNL
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

         362        372        382        392        402        412
Coq6_WI DRYCTDRVAL VGDAAHTTHP LAGQGLNMGQ TDVHGLVYAL EKAMERGLDI GSSLSLEPFW
2X3N    DRYWADNVAM LGDAIHNVHP ITGQGMNLAI EDASALADAL DLALRDACA- -LEDALAGYQ
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

         422        432        442        452        462        472
Coq6_WI AERYPSNNVL LGMADKLFKL YHTNFPPVVA LRTFGLNLTN KIGPVKNMII DTLGGNEK--
2X3N    AERFPVNQAI VSYGH----- ---------- ---------- ---------- ----------
4N9X    ---------- -----GFREL FDGDNPAKKL LRDVGLVLAD KLPGIKPTLV RQAXGLHDLP
```

FIG A5.2 *Construction alignment of Gen2 WT WI 2X3N_4N9X*

```
         25         35         45         55         65         75
Coq6_WI PKLTDVLIVG GGPAGLTLAA SIKNSPQLKD LKTTLVDMVD LKDKLSDFYN SPPDYFTNRI
4K22    MQSVDVAIVG GGMVGLAVAC GLQ----GSG LRVAVLEQR- ---------- -PPQ---LRV
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

         85         95        104        114        124         130
Coq6_WI VSVTPRSIHF LEN-NAGATL MHDRIQSYDG LYVTDGCSKA TLDLAR---- -DSMLCMIEI
4K22    SAINAASEKL LTRLGVWQDI LSRRASCYHG MEVWDKDSFG HISFDDQSMG YSHLGHIVEN
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

        139        149        158        168        178        188
Coq6_WI INIQASLYNR ISQYDSKKD- SIDIIDNTKV VNIKHSDPND PLSWPLVTLS NGEVYKTRLL
4K22    SVIHYALWNK AHQS-----S DITLLAPAEL QQVAWGE--- --NETFLTLK DGSMLTARLV
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

        198        208        218        228        238        248
Coq6_WI VGADGFNSPT RRFSQIPSRG WMYNAYGVVA SMKLEYPPFK LRGWQRFLPT GPIAHLPMP-
4K22    IGADGANSWL RNKADIPLTF WDYQHHALVA TIRTEEP-HD AVARQVFHGE GILAFLPLSD
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

        257        267        277        287        344        354
Coq6_WI ENNATLVWSS SERLSRLLLS LPPESFTALI NAAFVLEDA- VSIIDKTRAR FPLKLTHADR
4K22    PHLCSIVWSL SPEEAQRMQQ ASEDEFNRAL NIAFDNRLGL CKV-ESARQV FPLTGRYARQ
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

        364        374        384        394        404        414
Coq6_WI YCTDRVALVG DAAHTTHPLA GQGLNMGQTD VHGLVYALEK AMERGLDIGS SLSLEPFWAE
4K22    FASHRLALVG DAAHTIHPLA GQGVNLGFMD AAELIAELKR LHRQGKDIGQ YIYLRRYERS
1PBE    ---------- ---------- ---------- ---------- ---------- ----------

        424        434        444        454        464        474
Coq6_WI RYPSNNVLLG MADKLFKLYH TNFPPVVALR TFGLNLTNKI GPVKNMIIDT LGGNEK----
4K22    RKH------- ---------- ---------- ---------- ---------- ----------
1PBE    ---RIWKAER FSWWMTSVLH RFPDTDAFSQ RIQQTELEYY LGSEAGLATI AENYVGLPYE
```

FIG A5.3 *Construction alignment of Gen2 WT WI 4K22_1PBE*

```
         25         35         45         55         65         75
Coq6_WI PKLTDVLIVG GGPAGLTLAA SIKNSPQLKD LKTTLVDMVD LKDKLSDFYN SPPDYFTNRI
4K22    MQSVDVAIVG GGMVGLAVAC GLQ----GSG LRVAVLEQR- ---------- -PPQ---LRV
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

         85         95        104        114        124         130
Coq6_WI VSVTPRSIHF LEN-NAGATL MHDRIQSYDG LYVTDGCSKA TLDLAR---- -DSMLCMIEI
4K22    SAINAASEKL LTRLGVWQDI LSRRASCYHG MEVWDKDSFG HISFDDQSMG YSHLGHIVEN
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

        139        149        158        168        178        188
Coq6_WI INIQASLYNR ISQYDSKKD- SIDIIDNTKV VNIKHSDPND PLSWPLVTLS NGEVYKTRLL
4K22    SVIHYALWNK AHQS-----S DITLLAPAEL QQVAWGE--- --NETFLTLK DGSMLTARLV
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

        198        208        218        228        238        248
Coq6_WI VGADGFNSPT RRFSQIPSRG WMYNAYGVVA SMKLEYPPFK LRGWQRFLPT GPIAHLPMP-
4K22    IGADGANSWL RNKADIPLTF WDYQHHALVA TIRTEEP-HD AVARQVFHGE GILAFLPLSD
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

        257        267        277        287        344        354
Coq6_WI ENNATLVWSS SERLSRLLLS LPPESFTALI NAAFVLEDA- VSIIDKTRAR FPLKLTHADR
4K22    PHLCSIVWSL SPEEAQRMQQ ASEDEFNRAL NIAFDNRLGL CKV-ESARQV FPLTGRYARQ
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

        364        374        384        394        404        414
Coq6_WI YCTDRVALVG DAAHTTHPLA GQGLNMGQTD VHGLVYALEK AMERGLDIGS SLSLEPFWAE
4K22    FASHRLALVG DAAHTIHPLA GQGVNLGFMD AAELIAELKR LHRQGKDIGQ YIYLRRYERS
4N9X    ---------- ---------- ---------- ---------- ---------- ----------

        424        434        444        454        464        474
Coq6_WI RYPSNNVLLG MADKLFKLYH TNFPPVVALR TFGLNLTNKI GPVKNMIIDT LGGNEK----
4K22    RKH------- ---------- ---------- ---------- ---------- ----------
4N9X    ---SAAVXLA SXQGFRELFD GDNPAKKLLR DVGLVLADKL PGIKPTLVRQ AXGLHDLPDW
```

FIG A5.4 *Construction alignment of Gen2 WT WI 4K22_4N9X*

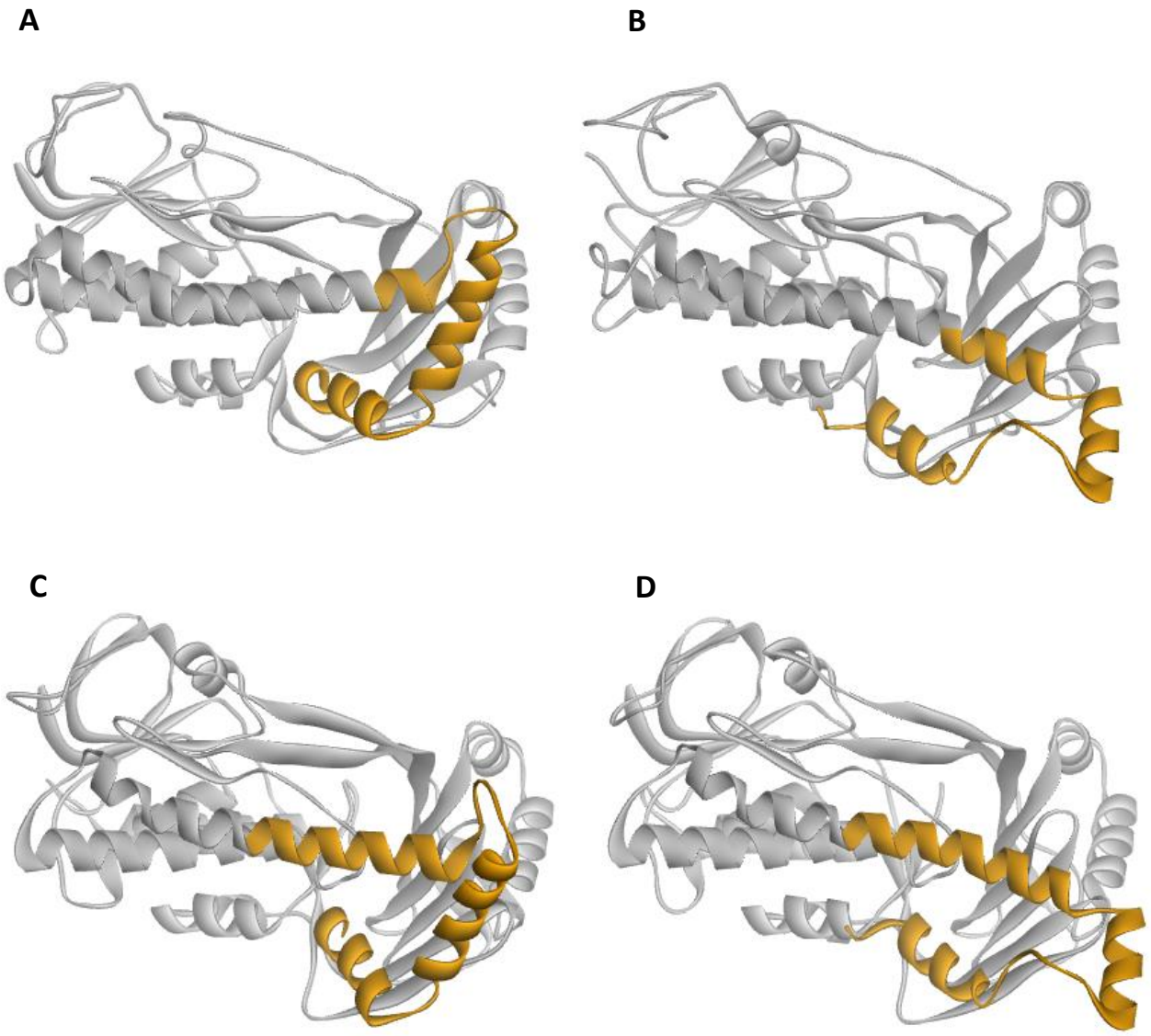

FIG A5.5 *Presentation of the multiple-template Gen2 Coq6 WT WI models, backbone trace only. Grey: N-terminus, orange, C-terminus. Models are named after their coordinate sources for N-terminus and C-terminus respectively. A) Based on 2X3N_1PBE, B) Based on 2X3N_4N9X, C) Based on 4K22_1PBE, D) Based on 4K22_4N9X.*

Each of the models faithfully inherits the geometry from each regional template. The models with the N-terminal region modeled on 4K22 coordinates have reproduced the distorted GDAxH motif in the FAD binding region, In the 2X3N based models this region adopts its FAD-binding compatible formation. The C-termini of the model reproduce their templated geometry. In the 1PBE based models the C-terminus is composed of three helical segments arranged into a roughly equilateral triangle. In the 4N9X based models, the C-terminus is also triangular, but the helical segments are shorter with longer turn regions between them, and the triangle is in a more extended conformation. To distinguish between these geometries, we will now subject these models to molecular dynamics.

3. MD simulation of Generation 2 WI constructs

The Gen2 models without the insertion were subjected to molecular dynamics simulations in order to assess their structural stability. The protein models were solvated using TIP3P water under the AMBER99SB-ILDN force-field using PME electrostatics in GROMACS 4.6.5. No FAD was included in this round of simulation for the Gen2 without-insert constructs. This is because some of the templates are experimentally known not to bind FAD, and we did not want this variable to a complicating factor in the interpretation of the simulations. The salt concentration was set to 0.157M NaCl as documented for the mitochondrial matrix. The simulation cell was a rhombic dodecahedron allowing 1.4 nanometers between the protein and the box edge. Models were subjected to 300 000 steps of steepest descent minimization. Equilibration was conducted in two phases (NVT and NPT) of 250 ps each at a time step of 1fs with position restraints on protein heavy atoms using the velocity-rescaled Berendsen thermostat at 300K. NPT equilibration used the Parinello-Rahman barostat. Bond lengths were not constrained during equilibration.

The Gen2 set of models tested two possibilities for the N-terminal template (4K22 and 2X3N) and two possibilities for the C-terminal template (1PBE and 4N9X). The purpose of the molecular dynamics screening of this panel is to evaluate the structural stability of each model, as stability is a necessary but not sufficient pre-requisite for functionality (which is explored later through other calculations). This is an important distinction because models based on catalytically inactive templates may still be dynamically stable. This distinction must be kept in mind primarily for the N-terminal region of the models because it comprises the majority of the protein and has many well formed elements of secondary and super-secondary structure likely to remain stable despite lack of catalytic activity. However, the C-terminal region of the models is much smaller and has many fewer long- range contacts with the rest of the protein, making it more dependent on being properly structured intrinsically in order to be stable. Therefore, when reviewing the trajectories of the Gen2 WI models we will focus on the behavior of the C-terminus. The structural stability of the C-terminus can be described by how well it retains its secondary structure. Secondary structure can be calculated on the basis of the protein's internal coordinates (the backbone angles *phi* and *psi)* and it is a more accurate and robust description of local structural stability than atomic RMSD. This is because of the inherently non-directional averaging of atomic coordinates relative to the reference coordinates in the final computed RMSD value.

Two extreme cases of RMSD being an inaccurate descriptor of local structural stability exist. One is the rigid body displacement of a structure from its initial position. The conformation of the structure itself may be perfectly rigid, but an RMSD curve calculated based on its initial position will show a continuous increase, implying a structural deviation where there has only been a positional one. The other case is a conformational denaturation distributed evenly over the entire initial structure held at a fixed center of mass. In this case the RMSD curve calculated over time will be relatively flat, implying structural stability, while the conformation has changed significantly.

Therefore, to give a synoptic review of the MD simulations for this set we will show the first and last frames from the trajectory as well as the secondary structure description of the C-terminal region computed over the duration of the trajectories.

3.1 Gen2 Without Insertion 4K22 – 1PBE

In this model the N-terminus has been modeled after 4K22. The C-terminus has been modeled after 1PBE, (shown in orange in the figure below) giving it a large equilateral triangle shape composed of three helical segments. However, this conformation is not stable. This is visible from the difference between the first (A) and last (B) frames of dynamics, which show that the last helix unwinds, as does the penultimate helix. The helical segment preceding this also undergoes a major conformational change, becoming irregular and tilted upwards. This is also described in the secondary structure plot below (C).

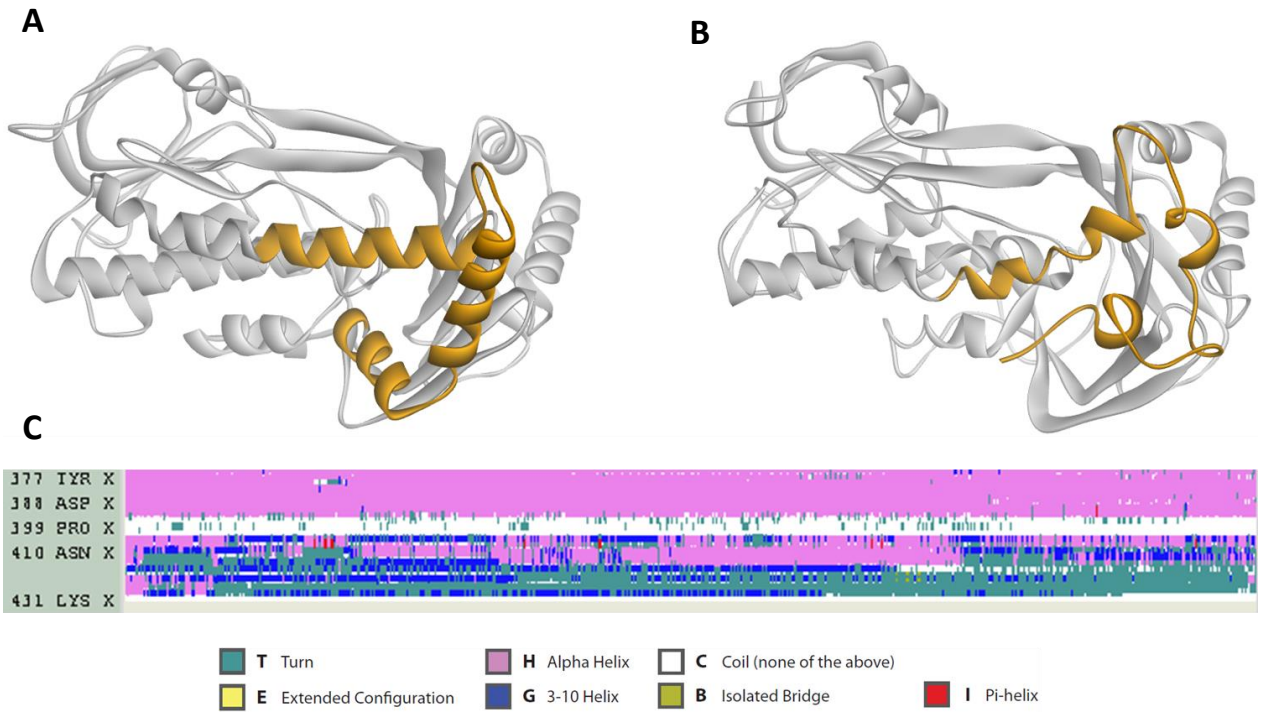

FIG A5.6 *Generation 2 4K22-1PBE based Coq6 homology model: molecular dynamics stability screening review. A) The model before dynamics, B) After 20ns dynamics, C) secondary structure persistence plot as calculated by the Timeline plugin for VMD.*

3.2 Gen2 Without Insertion 4K22 – 4N9X

In this model the N-terminus has been modeled after 4K22. The C-terminus has been modeled after 4N9X, (shown in orange in the figure below), also giving it a generally triangular conformation. However, the three helical segments composing it are shorter, resulting in a more extended, less equilateral triangle. This conformation is not stable. This is visible from the difference between the first (A) and last (B) frames of dynamics, which show that the last and penultimate helices unwind. Most significantly, a portion of the long N-terminal helical segment breaks abruptly, displacing the entire C-terminus downwards. This change, while less noticeable in the secondary structure plot, is easily visible in 3D.

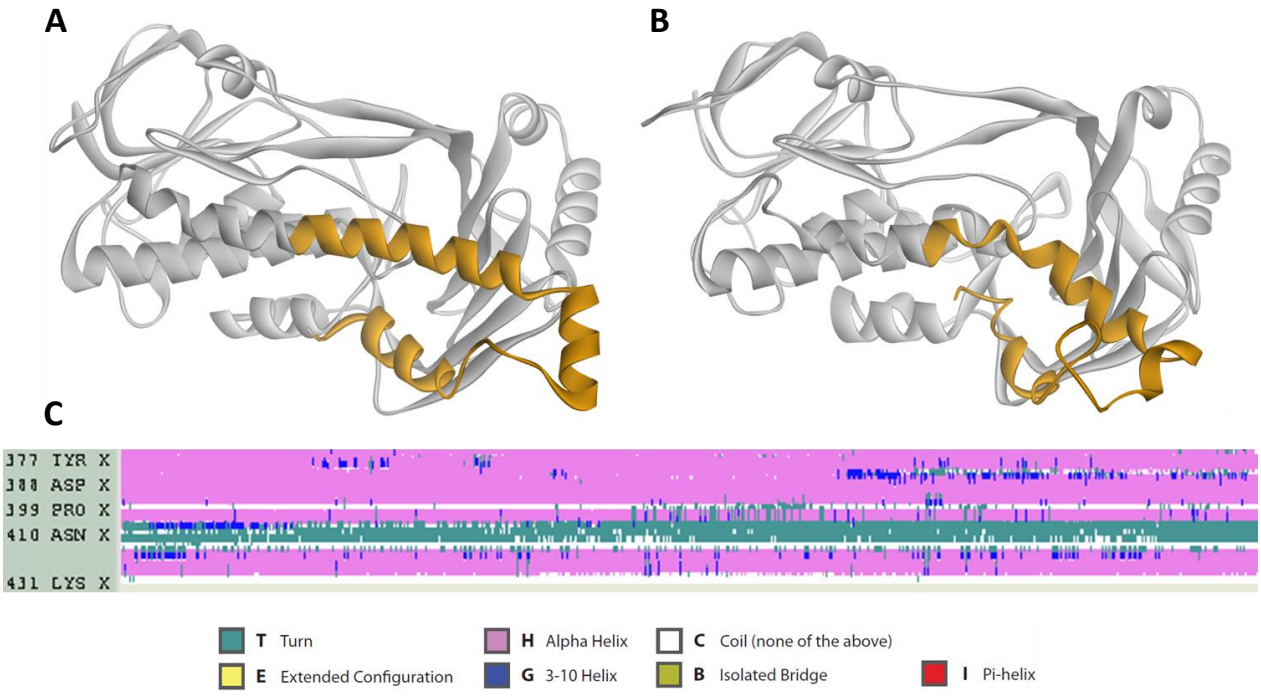

FIG A5.7 *Generation 2 4K22-4N9X based Coq6 homology model: molecular dynamics stability screening review. A) The model before dynamics, B) After 20ns dynamics, C) secondary structure persistence plot as calculated by the Timeline plugin for VMD.*

3.3 Gen2 Without Insertion 2X3N – 1PBE

In this model the N-terminus has been modeled after 2X3N. The C-terminus has been modeled after 1PBE, (shown in orange in the figure below). The initial conformation remains a large, roughly equilateral triangle. The last two helices remain helical, but they re-organize into a different tertiary structure by changing geometry at the intervening turn regions, aligning into an anti-parallel helical bundle. This indicates that the C-terminal structure inherited from 1PBE is not compatible with a core modeled after 2X3N – the only C-terminal template with a chance of being catalytically active.

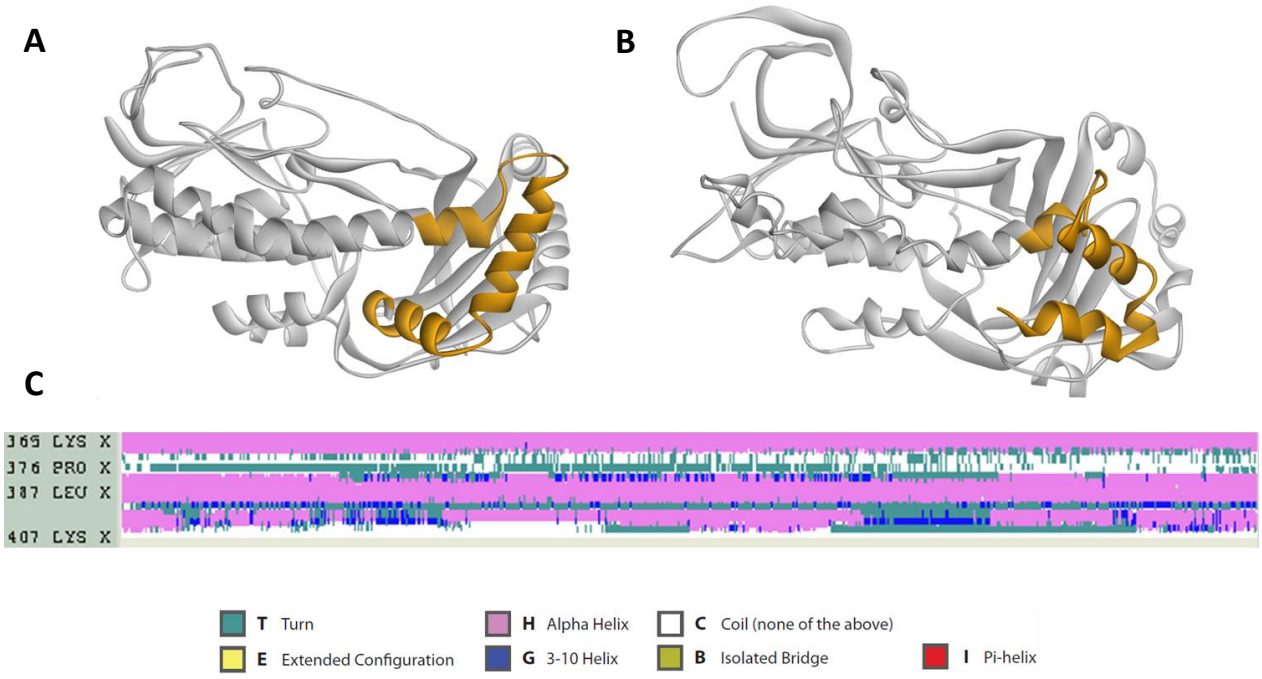

FIG A5.8 *Generation 2 2X3N-4N9X based Coq6 homology model: molecular dynamics stability screening review. A) The model before dynamics, B) After 20ns dynamics, C) secondary structure persistence plot as calculated by the Timeline plugin for VMD.*

3.4 Gen2 Without Insertion 2X3N – 4N9X

In this model the N-terminus has been modeled after 2X3N. The C-terminus has been modeled after 4N9X, (shown in orange in the figure below). The initial C-terminal conformation is inherited from 4N9X as an extended triangle composed of three helical segments. Dynamics reveals an interesting behavior: the penultimate helix starts in a vertical orientation, as in panel A below, but can rotate to a more horizontal conformation, as in panel B below, while maintaining its helical content, moving primarily as a rigid body. The preceding and succeeding helical segments remain well structured. The stability of the C-terminus is also readily apparent from the secondary structure persistence plot in panel C, which is dominated by continuous horizontal streaks of purple (corresponding to alpha helical secondary structure with very few interruptions.

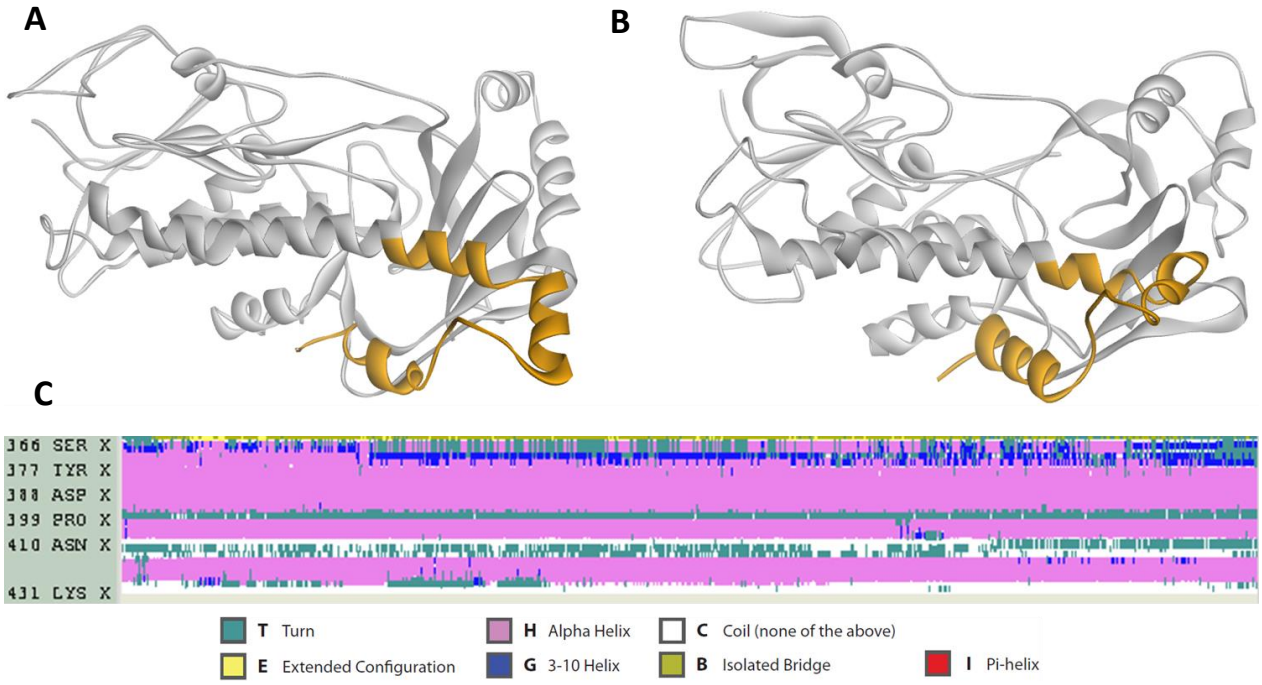

FIG A5.9 *Generation 2 4K22-4N9X based Coq6 homology model: molecular dynamics stability screening review. A) The model before dynamics, B) After 20ns dynamics, C) secondary structure persistence plot as calculated by the Timeline plugin for VMD.*

4. Conclusion

Of the Generation 2 WI (without insertion) Coq6 models tested here, we conclude that only the 2X3N-4N9X based model presents a structure stable over 20ns of molecular dynamics. Fortunately, the 2X3N based moiety is also compatible with FAD binding for catalysis, since the original 2X3N crystal structure contains it. Therefore, we select the 2X3N and 4N9X structures as templates for generating models of Coq6. The next phase will be to integrate the Coq6-family insert to the model. In addition, the precise sequence definition of the Coq6 insert can be used by our partner, Dr. Pierrel, to suggest excisions and inter-species complementation assays to test its function.

Annex 6 FAD parameter values

Because of space considerations, the parameter table is presented in a separate electronic file.

Annex 7 Journal article

Article as published in the Journal of Biological Chemistry in 2013

Because of space considerations, this article is presented in a separate electronic file.

ubiI, a new gene in _Escherichia coli_ coenzyme Q biosynthesis, is involved in aerobic C5-hydroxylation

Chehade, Mahmoud Hajj, Laurent Loiseau, Murielle Lombard, Ludovic Pecqueur, Alexandre Ismail, Myriam Smadja, Béatrice Golinelli-Pimpaneau, et al. "ubiI, a New Gene in Escherichia Coli Coenzyme Q Biosynthesis, Is Involved in Aerobic C5-Hydroxylation." Journal of Biological Chemistry 288, no. 27 (July 5, 2013): 20085–92. doi:10.1074/jbc.M113.480368.

Annex 8 Journal article, first author

Manuscript as submitted to PLOS Computational Biology in 2015

Because of space considerations, this article is presented in a separate electronic file.

Coenzyme Q biosynthesis: Evidence for a substrate access channel in the FAD-dependent monooxygenase Coq6

Ismail, Alexandre, Vincent Leroux, Myriam Smadja, Lucie Gonzalez, Murielle Lombard, Fabien Pierrel, Caroline Mellot-Draznieks, and Marc Fontecave. "Coenzyme Q Biosynthesis: Evidence for a Substrate Access Channel in the FAD-Dependent Monooxygenase Coq6." *PLoS Comput Biol* 12, no. 1 (January 25, 2016): e1004690. doi:10.1371/journal.pcbi.1004690.

Annex 9 Journal article

Article as published in the Journal of Biological Chemistry in 2015

Because of space considerations, the results for each Coq protein are presented in a separate electronic file.

Coq6 is responsible for the C4-deamination reaction in coenzyme Q biosynthesis in _Saccharomyces cerevisiae_

Ozeir, Mohammad, Ludovic Pelosi, Alexandre Ismail, Caroline Mellot-Draznieks, Marc Fontecave, and Fabien Pierrel. "Coq6 Is Responsible for the C4-Deamination Reaction in Coenzyme Q Biosynthesis in Saccharomyces Cerevisiae." Journal of Biological Chemistry, August 10, 2015, jbc.M115.675744. doi:10.1074/jbc.M115.675744.