Strategies for rational protein engineering of cytochrome P450 monooxygenase systems

Von der Fakultät Energie-, Verfahrens- und Biotechnik

der Universität Stuttgart zur Erlangung der Würde eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigte Abhandlung

Vorgelegt von

Łukasz Gricman

geboren in Olsztyn, Polen

Hauptberichter: Prof. Dr. Jürgen Pleiss

Mitberichter: Prof. Dr. Ralf Takors

Tag der mündlichen Prüfung: 12.10.2015

Institut für Technische Biochemie der Universität Stuttgart

2015

Acknowledgments

This dissertation would not be possible without numerous collaborations and support received from colleagues, friends and family.

I am especially grateful to Prof. Dr. Jürgen Pleiss for giving me the opportunity to work on this project. I would like to express my appreciation for the supervision, motivation and freedom to explore my interests.

I would like to thank Prof. Dr. Bernhard Hauer for the opportunity to conduct my research at the Institute for Technical Biochemistry.

I would like to thank Prof. Dr. Ralf Takors for the time and effort spend on reviewing this dissertation.

I would like to thank European Union for the financial support in the framework of Marie Curie Actions Initial Training Network P4FIFTY under REA Grant Agreement 289217 and High Performance Computing Centre Stuttgart (HLRS) and the BW-UniCluster for access to computational resources

I would like to thank all the members of the P4FIFTY network for helpful discussions and stimulating atmosphere at the biannual meetings. My special thanks go to the network members with whom I had pleasure to collaborate on interesting projects: Anja Eichler and Prof. Dr. Sabine Flitsch, Ilona Jóźwik and Prof. Dr. Andy-Mark Thunnissen, Nina Beyer and Prof. Dr. Dick Janssen, Dr. Michael Ringle and Dr. Jürgen Riegler.

I would like to thank all members of the Institute for Technical Biochemistry. In particular many thanks go to my collaboration partners Dr. Constantin Vogel, Dr. Martin Weissenborn, Sandra Notonier, Sara Hoffman and Niels Borlinghaus.

I am grateful to the Bioinformatics group for their help and companionship over the last three years, without Conny, Silvia, Sven, Quy and Waldemar my time at the ITB would be much less fun. Special thanks go to Conny who has critically read the dissertation and helped me with translation of the abstract.

Finally, I would like to thank Agnieszka, family and friends for their continuous love and support.

Contents of this dissertation have been already published or submitted for publication:

I. **Gricman Ł.**, Vogel C., Pleiss J., 2014. Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance. Proteins 82: 491-504

II. **Gricman Ł.**, Vogel C., Pleiss J., 2015. Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining. Proteins 83:1593-1603

III. **Gricman Ł.**, Weissenborn M., Hoffmann S., Borlinghaus N., Hauer B., Pleiss J., 2015. Redox partner interaction sites in cytochrome P450 monooxygenases: *in silico* analysis and experimental validation. *Manuscript in preparation*

TABLE OF CONTENTS

Abstract

Cytochrome P450 monooxygenases (CYPs) constitute a large and diverse protein family. Because of their ability to introduce molecular oxygen into non-activated C-H bonds, CYPs are interesting enzymes for synthetic applications. However, stability, selectivity and activity are often obstacles prohibiting those enzymes from being used in industrial processes. Thus, novel protein engineering strategies for overcoming all those bottlenecks are required. In this work, a comprehensive guide to protein engineering of CYPs covering those three aspects was established to aid rational protein engineering of enzymes with and without known structure. The described protein engineering strategies are based on new insights into CYPs, which were gathered from incorporation of new functionalities into and analyses of the Cytochrome P450 Engineering Database (www.CYPED.BioCatNet.de), as well as molecular modeling of stereoselectivity in CYP101A1 from *Pseudomonas putida*. The described strategy for improving protein stability is based on the insights gathered from a systematic sequence-based protein stabilization of bovine adrenodoxin reductase. The presented strategies for protein engineering of selectivity are based on the identification of universal selectivity determining positions by systematic literature mining and on new insights from molecular modeling of stereoselectivity. The strategies for protein engineering of activity are based on the amino acid conservation analysis of CYPs, identification of the redox partner interaction sites, and a strategy for re-designing linker regions in artificial CYP fusion systems.

Zusammenfassung

Die Cytochrom P450-Monooxygenasen formen eine große und diverse Proteinfamilie. Aufgrund ihrer Fähigkeit molekularen Sauerstoff in nicht-aktivierte C-H-Bindungen einzufügen, sind Vertreter dieser Familie von zunehmendem industriellem Interesse. Ihre Anwendbarkeit in nicht-physiologischem Umfeld wird jedoch häufig durch mangelnde Stabilität und ungeeignete Selektivitäten und Aktivitäten eingeschränkt. Zur Überwindung dieser Einschränkungen besteht daher Bedarf an neuartigen Strategien des Protein-Engineerings. Im Rahmen dieser Arbeit wurde ein umfangreicher Leitfaden zur Anpassung von Stabilität, Selektivität und Aktivität von Cytochrom P450-Monooxygenasen an die Anforderungen synthetischer Anwendungsfelder entworfen. Dieser Leitfaden enthält Strategien zur Optimierung von Enzymen mit bekannter wie auch mit unbekannter Struktur. Die vorgestellten Strategien beruhen auf neuen Erkenntnissen über Cytochrom P450-Monooxygenasen, die in dieser Arbeit durch Einsatz verschiedener Methoden gewonnen werden konnten: durch Einbindung neuer Funktionen in die Cytochrome P450 Engineering Database (www.CYPED.BioCatNet.de), deren Aktualisierung und Analyse, sowie durch molekulare Modellierung der Stereoselektivität in der CYP101A1 aus *Pseudomonas putida*. Ausgehend von einer systematischen, sequenz-basierten Stabilisierung der Rinder-Adrenodoxin Reduktase beschreibt diese Arbeit eine Strategie zur Verbesserung der Proteinstabilität in Cytochrom P450-Monooxygenase-Komplexen. Die hier beschriebene Strategie zur Modifikation der Proteinselektivitäten beruht auf der Identifikation von universellen selektivitätsbestimmenden Positionen durch systematische Auswertung wissenschaftlicher Literatur und neuen Erkenntnissen durch molekulare Modellierung der Stereoselektivität in P450-Monooxygenasen. Zur Verbesserung der Enzymaktivität beschreibt der Leitfaden eine kombinierte Strategie unter Verwendung einer Analyse der Aminosäure-Konservierung in Cytochrom P450-Monooxygenasen, der Identifikation von Redoxpartner-Interaktionsstellen und der Umgestaltung von Linker-Regionen in artifiziellen CYP-Fusionssystemen.

1. Introduction

This chapter consists of information about the mechanism, function and classification of cytochrome P450 monooxygenases. Furthermore, major protein engineering methods are introduced in light of the results presented in this dissertation.

1.1 Cytochrome P450 monooxygenases

Cytochrome P450 monooxygenases (CYPs) are a diverse protein family that can be found in all domains of life. The name of CYPs is tightly bound to the way those enzymes were discovered. In the late 1950s', Axelrod and Brodie identified a red (pigment), liver endoplasmic reticulum enzyme capable of performing oxidation of xenobiotic compounds, therefore the enzyme was classified as a monooxygenase.[1,2] Subsequently, it was discovered that the enzyme binds CO and has an absorption maximum at 450 nm, afterwards it was characterized as a hemoprotein (cytochrome). Hence, it was named cytochrome P450 – the 'P' stands for pigment.[3–6]

All CYPs contain a catalytically active heme cofactor covalently bound to a conserved cysteine by a Fe-S bond, which is responsible for its spectroscopic properties and the Soret peak[a] at 450 nm.[7] CYPs play diverse roles in living organisms, from biotransformation of xenobiotic compounds to being part of biosynthetic pathways in metabolite synthesis.[8,9] Because of their importance in human health and interesting catalytic reactions CYPs are of industrial interest.[10,11]

1.1.1 Nomenclature of CYPs

Nomenclature of CYPs was established to simplify communication about members of the protein family. Since, it is based on a combination of the taxonomy of source organism and the sequence identity between family members it also provides an idea about the enzymes'

[a] Soret peak – in spectroscopy refers to a wavelength of maximum adsorption in the blue region of visible spectrum at around 400 nm. CYPs in their reduced form when saturated with CO exhibit a Soret peak at 450 nm, in denatured CYPs the peak shifts to 420 nm.

origin and similarity to other CYPs. The naming scheme was first introduced in late 1980s', and since then was updated to facilitate increasing number of identified CYPs.[12–16] The sequence identity criterion divides the family into gene families (in this dissertation called superfamilies) and subfamilies (in this dissertation called homologous families). Proteins with more than 55% sequence identity are grouped into homologous families, which are organized into superfamilies based on 40% identity cut-off. An example of a CYP name is CYP1A1, where the first '1' stands for the superfamily 1, 'A' stands for homologous family A, and '1' means that it is the first protein identified in this homologous family. To include taxonomic classification of CYP variants into the nomenclature a set of superfamily number intervals was introduced into the naming scheme. Thus, enzymes of animal origin are assigned to superfamilies CYP1-50, CYP301-499 and CYP3001-4999, lower eukaryotes to CYP51-70, CYP501-699 and CYP5001-6999, plants to CYP71-100, CYP701-999 and CYP7001-9999, and prokaryotes to CYP101-299 and CYP1001-2999. Naming of the new CYPs is done by David R. Nelson. As of the last update in August 2013 there were 21039 CYPs named on the Cytochrome P450 Homepage (www.drnelson.uthsc.edu/CytochromeP450.html). The current version of Cytochrome P450 Engineering Database established in November 2015 (ww.CYPED.BioCatNet.de) contains sequences of 42000 CYPs. With the exponentially growing number of sequences it might be necessary to design a centralized and automated way of assigning CYP names. An example of such a tool would be a webserver accepting sequences uploaded by researchers, which would then be checked against a database of already identified CYPs and subsequently named accordingly to taxonomy of the source organism and location in the sequence space.

1.1.2 Catalytic mechanism and reactions

The catalytic mechanism of cytochrome P450 monooxygenases was initially introduced in 1968.[17] This first description is a foundation of our current understanding of the CYPs catalytic cycle. Over the last half a century the mechanism was extended to accommodate newly discovered reaction pathways and intermediates.[18–20]

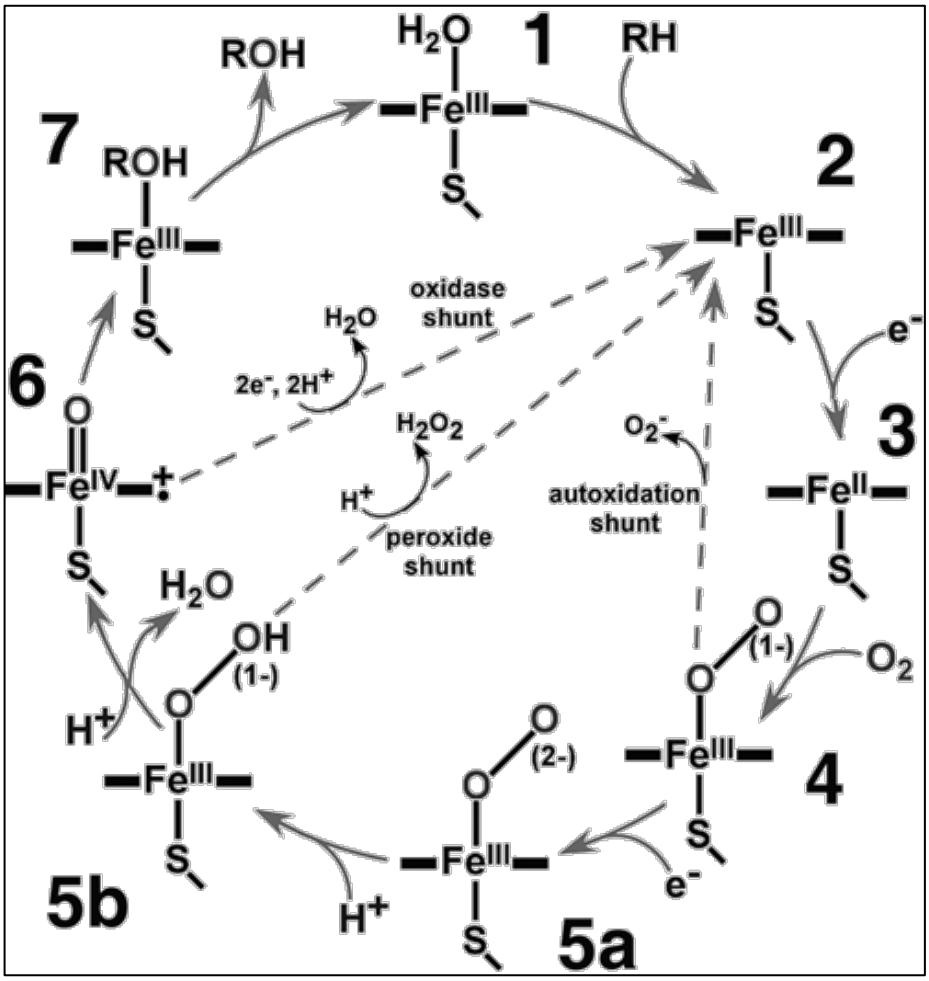


Figure 1: Catalytic cycle of CYPs. The reaction pathway is shown by arrows between the schematic representations of heme intermediates. Side reactions are shown as the broken lines. (Adapted from Denisov et al. 2005).[20]

The catalytic cycle of CYPs can be described in seven consecutive steps (Figure 1). At the beginning of the cycle, upon entrance of the substrate into the active site, the resting state heme iron bound water is displaced (1, 2). The loss of water coordination of the heme causes a spin shift of the heme iron from a low-spin to a high-spin state, which in turn allows for reduction of the iron by the first electron delivered from the redox partner (3). Subsequently the iron binds oxygen and forms an oxy-ferrous complex (4). The second electron transfer reduces the oxy-ferrous complex and produces a negatively charged iron-peroxo complex, which is then protonated and forms an iron-hydroxyperoxo complex (5a and 5b). Subsequent protonation event allows for formation of the reactive iron-oxo complex also referred to as compound I, and release of water (6). The final step of the reaction is oxidation of the substrate by the reactive compound I and release of the product (7). The cycle can be interrupted, which in turn may result in so called "electron uncoupling", meaning that not all

electrons delivered to the heme iron are contributing to product formation. There are three uncoupling pathways: autoxidation shunt, peroxide shunt and oxidase shunt. The autoxidation shunt takes place when transfer of the second electron is disturbed. The peroxide shunt is caused by the disintegration of the iron-hydroxyperoxo complex and release of hydrogen peroxide. The oxidase shunt appears when the iron-oxo complex does not oxidate the substrate but water instead. The uncoupling reactions do not only decrease the turnover number but can also damage the enzyme and its host cell by production of hydrogen peroxide.[21–23]

Cytochrome P450 monooxygenase catalyze a slew of different reactions by insertion of single oxygen atom into the substrate. CYP catalyzed reactions include aliphatic hydroxylation, aromatic hydroxylation, N- and O-dealkylation, deamination, desulfurization, etc..[24,25] CYPs have been successfully applied in industrial processes, and have high potential for synthetic applications.[10]

1.1.3 Classification of CYPs by redox partner type

Cytochrome P450 monooxygenases require two electrons for their catalytic activity, in most of the CYPs the electrons are delivered to the heme by an external redox partner. Based on the natively associated redox partners the family can be divided into at least ten classes (Figure 2 on the following page).[26]

Class I CYPs (Figure 2A) are mostly of prokaryotic origin, and constitute the biggest group amongst known prokaryotic CYPs. Exception in this class are mitochondrial CYPs (Figure 2B), which differ from the bacterial class I enzymes by the fact that the CYP and ferredoxin reductase are respectively membrane bound and membrane associated. Class I CYPs natively accept electrons from a ferredoxin (iron-sulfur cluster protein), which is shuttling them from NAD(P)H dependent ferredoxin reductase. The most prominent member of this class is CYP101A1 from *Pseudomonas putida* (P450cam), which is one of the most studied CYPs and model system for studying the reaction mechanism. **Class II** CYPs (Figure 2C) constitute the biggest group amongst eukaryotic CYPs. In this class electrons are delivered to the heme by the two domain diflavin cytochrome P450 reductase (CPR). The reductase contains FAD and FMN cofactors and transfers the electrons from NADPH. In this class both CYP and CPR are membrane bound.

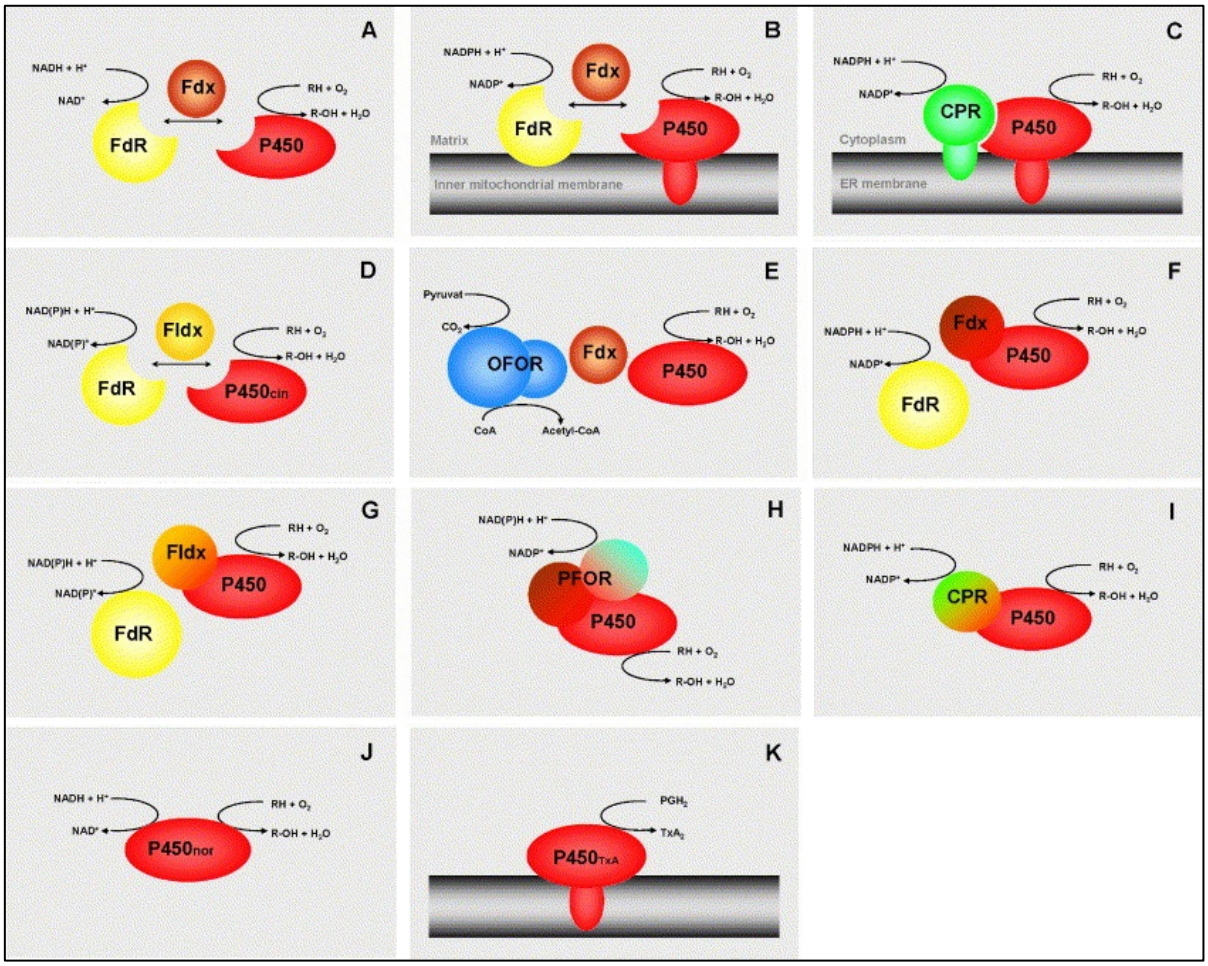


Figure 2: Schematic representation of different electron transfer chains in CYPs. (A) Class I, prokaryotic system; (B) class I, mitochondrial system; (C) class II microsomal system; (D) class III, prokaryotic system; example P450cin; (E) class IV, prokaryotic thermophilic system; (F) class V, prokaryotic [Fdx]–[P450] fusion system; (G) class VI, prokaryotic [Fldx]–[P450] fusion system; (H) class VII, prokaryotic [PFOR]–[P450] fusion system; (I) class VIII, prokaryotic [CPR]–[P450] fusion system; (J) class IX, soluble eukaryotic P450nor; (K) independent eukaryotic system, example P450TxA. (Adapted from Hannemann et al. 2007)[26]

Class III CYPs (Figure 2D) similarly to class I, are natively part of a three component electron transfer chain, but here the electrons are transferred from the NAD(P)H via the FAD containing flavodoxin reductase and the FMN containing flavodoxin (as compared to class I iron-sulfur cluster ferredoxin) to the CYP. Up to now only prokaryotic CYPs were described to belong to class III. **Class IV** (Figure 2E) was introduced after identification of the thermophilic CYP119 from *Sulfolobus solfataricus*,[27] it was the first characterized CYP which did not require a NADP(H)-dependent flavoprotein. In class IV the electrons are shuttled from a 2-oxoacid:ferredoxin oxidoreductase to the ferredoxin and then to the CYP. **Class V** systems (Figure 2F) similarly to class I consist of three domains: ferredoxin reductase, ferredoxin and CYP, but in this class the CYP and ferredoxin are one fusion protein. **Class VI**

systems (Figure 2G) contain a flavoprotein reductase and a fusion protein of a flavodoxin and a CYP. **Class VII** (Figure 2H) CYPs are fusion systems analogous to class I, combining CYP, ferredoxin and ferredoxin reductase into one protein. In this class electrons are transferred from NADH to a FMN cofactor and reach the heme via an iron-sulfur cluster domain. **Class VIII** (Figure 2I) contains other fusion protein systems. Those enzymes are fusion equivalents of class II where the CYP domain is fused with diflavin reductase component. The most prominent member of this class is the widely studied CYP102A1 from *Bacillus megaterium* (P450BM-3). **Class IX** (Figure 2J) up to now only contains CYP55 superfamily, and it is the only group of soluble eukaryotic CYPs. Enzymes from this class are independent from electron transfer proteins and use NADH directly as an electron donor. **Class X** (Figure 2K) is also a class of redox partner-independent CYPs. In this class, the reaction is catalyzed using intramolecular electron transfer systems.

For many newly identified CYPs the natural redox partners are not known, therefore artificial electron transfer chains are established.[28] Those electron transfer chains are often combined into **artificial fusion proteins** containing the redox partner domains of classes VII or VIII.[29,30] In the CYP systems combining non-natural redox partners, the electron transfer can be rate limiting factor, which is often caused by the non-optimal CYP-redox partner interactions. In chapters 3.3, 3.4 and 5.3, strategies aiming at improving those interactions by rational CYP-redox partner interface and linker re-design are described.

In this dissertation a simplified classification of CYPs will be used. The simplified classification is restricted to two classes. It is based on similarity of the smaller classes to class I and class II CYPs, which account for over 90% of all known family members.[31] The simplified class I contains members of the previously described classes I, III, IV, V, VI, and VII. All those CYPs require iron-sulfur cluster proteins or flavodoxins as electron donor partners and are in majority of prokaryotic origin. The simplified class II contains eukaryotic CYPs either self-sufficient or accepting electrons from CPR-type reductases and prokaryotic fusion systems containing the diflavin reductase. Self-sufficient CYPs were included in class II because of their mostly eukaryotic origin and structural similarity to other class II CYPs.

1.1.4 Sequence and structure diversity of CYPs

Cytochrome P450 monooxygenases share a common fold but as in many other protein families the sequence identity between the family members can be lower than 15%.[32,33] Even though the global sequence identity between CYPs is often low, there are certain positions which are conserved in the majority of those enzymes. The heme binding cysteine is the only fully conserved position in all CYPs. Another position important to the catalytic mechanism is a highly conserved threonine located on the α-helix I (Fig. 3).[34–36] There are few other positions and motifs described as conserved (e.g. the EXXR-motif[37]), but all of those positions were identified in studies covering only a limited section of the CYPs' sequence space. In the chapters 3.1 and 5.1 the first conservation analysis of all CYPs collected in the CYtochrome P450 Engineering Database (www.CYPED.BioCatNet.de) is described.[31] The analysis revealed highly conserved positions which were up to then not described to be conserved, and showed interesting differences in conservation of functionally relevant positions in class I and class II CYPs. The analysis was enabled by establishing two class-specific standard numbering schemes for CYPs, which allow for easy communication and comparison of structurally corresponding amino acid positions between different CYPs.

The first structure of a cytochrome P450 monooxygenase was the CYP101A1 from *Pseudomonas putida* (P450cam) with its natural substrate camphor in the binding pocket.[38] The structural elements constituting CYP-fold were named and numbered accordingly to this structure (Figure 3 on the following page).[38]

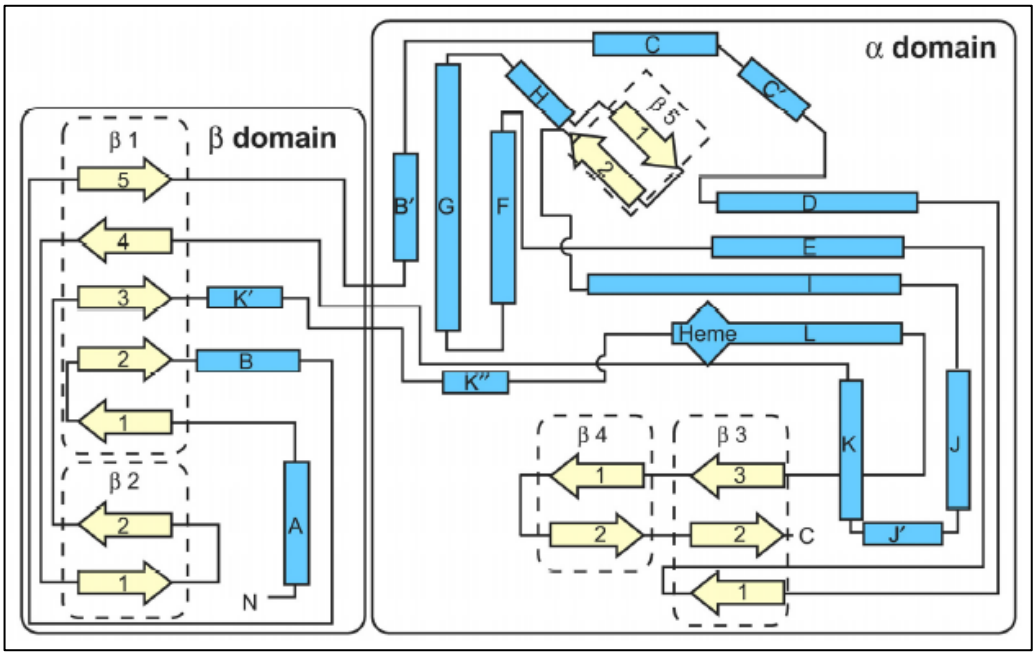


Figure 3: Schematic representation of the CYP-fold. α- and β-domains are separated by black outlines. α-helixes are shown as blue rectangles. β-strands are shown as pale yellow arrows, which are grouped into β-sheets by dotted outlines. (Adapted from Werck-Reinhart and Feyereisen 2000).[8]

The CYP-fold consists of thirteen conserved α–helices and five β–sheets, [38,39] the α–helices are alphabetically named A-L and the β–sheets are numbered 1-5. In this conserved fold six substrate recognition sites (SRSs) covering most of the substrate binding pocket and substrate entrance were identified (Figure 4 on the following page).[40] SRS1 is located on the β-strand 1-5-loop-α-helix B', SRS2 is located on α-helix F, SRS3 on α-helix G, SRS4 on α-helix I, SRS5 covers β-strand 1-4 and the neighboring loops, and SRS6 spans over β-strands 4-1 and 4-2. SRS2 and 3 constitute most of the substrate access channel, whereas SRS1, 4, 5, and 6 form the walls of the binding pocket. Because the SRS positions not only directly interact with the substrate but also contribute to the general architecture and flexibility of the binding pocket, it is clear that some SRS positions will be more significant for the determination of selectivity and specificity than other. Thus, with the sequence-based literature mining algorithm (SBLMA) described in the chapters 3.2 and 5.2, a new method developed and implemented into the CYPED to allow for identification of positions influencing selectivity among wide range of CYPs.

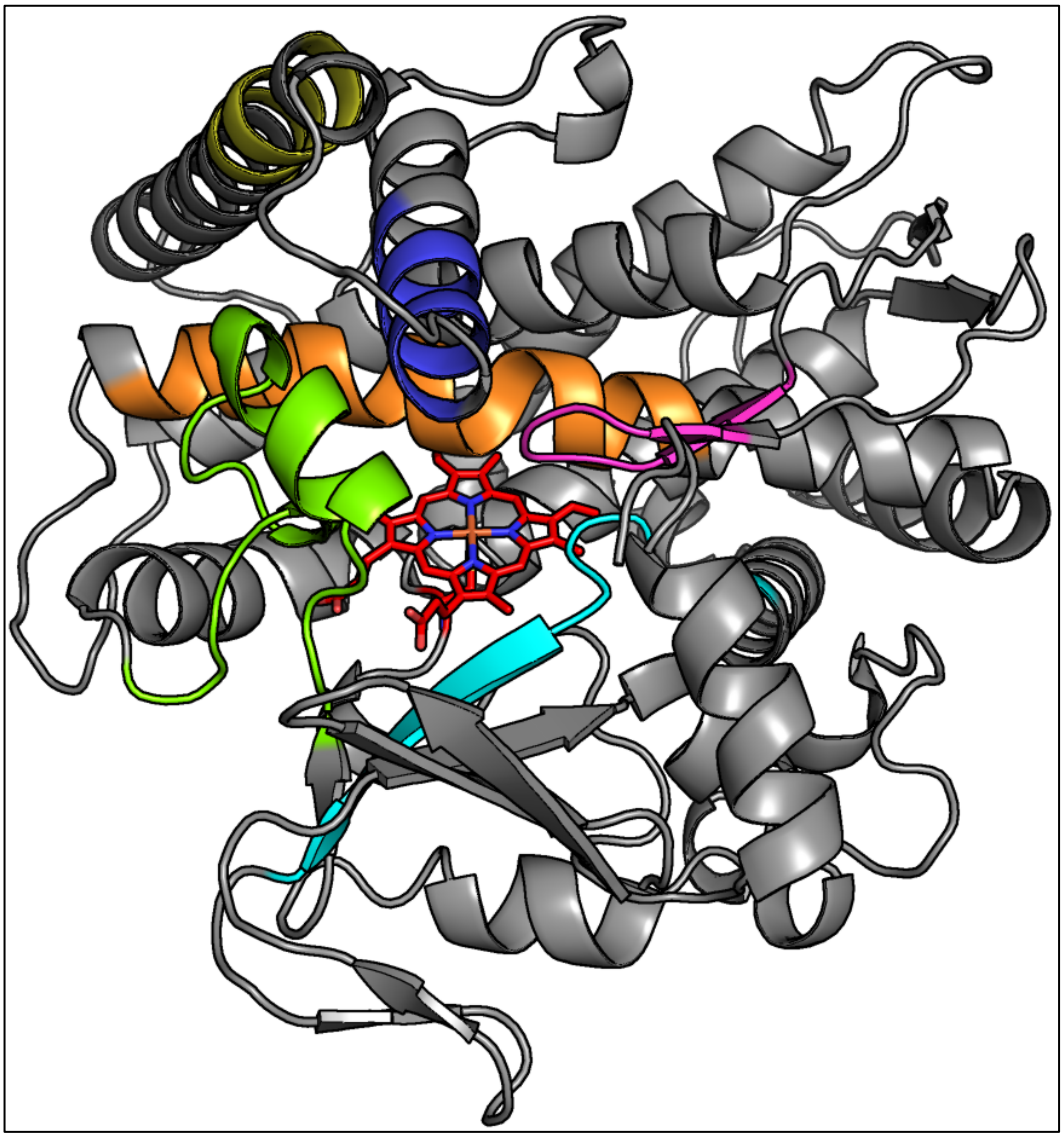


Figure 4: Substrate recognition sites on the structure of CYP102A1 (PDB: 1BVY). Heme is shown as red sticks, SRSs are highlighted in color, SRS1 green, SRS2 blue, SRS3 yellow, SRS4 orange, SRS5 cyan, SRS6 magenta.

In the current version of the CYPED (est. late 2014) 595 structures of 119 CYPs are stored. Majority of the structures come from prokaryotic organisms, this might be due to synthetic potential of those enzymes, but also more feasible crystallization process in comparison to membrane bound CYPs like the plant enzymes. There are crystal structures of plant CYPs only from the homologous family CYP74A which contains self-sufficient enzymes.[41,42] The difficulty with crystallization of plant CYPs might come from their membrane bound character and low solubility.[8] One of the other big challenges in CYP crystallography is obtaining structures of CYP-redox partner complexes. Up to now there are structures of three CYPs published with full or incomplete redox partners bound to a similar region of their proximal surfaces (Figure 5 on the following page). The first one was a structure of the

CYP102A1 heme domain with its reductase FMN domain (Figure 5A),[43] the only animal CYP with a partial redox partner is class I CYP11A1 with partial adrenodoxin bound to its proximal surface (Figure 5C).[44] The most recent crystal structure of CYP-redox partner complex was CYP101A1 with putidaredoxin, which was published in mid-2013 (Figure 5B).[45,46] The latter is probably the most reliable, because it contains complete structure of the redox partner and was reproduced by two independent groups using different methods. The structure of CYP101A1 with putidaredoxin confirms the proximal surface binding, which was already observed in the two previous crystals.

In the chapters 3.3 and 5.3 the current knowledge on the CYP-redox partner interactions is extended by identification of the redox partner interaction sites (RPIS), and is used in engineering of the CYP-redox partner interactions.

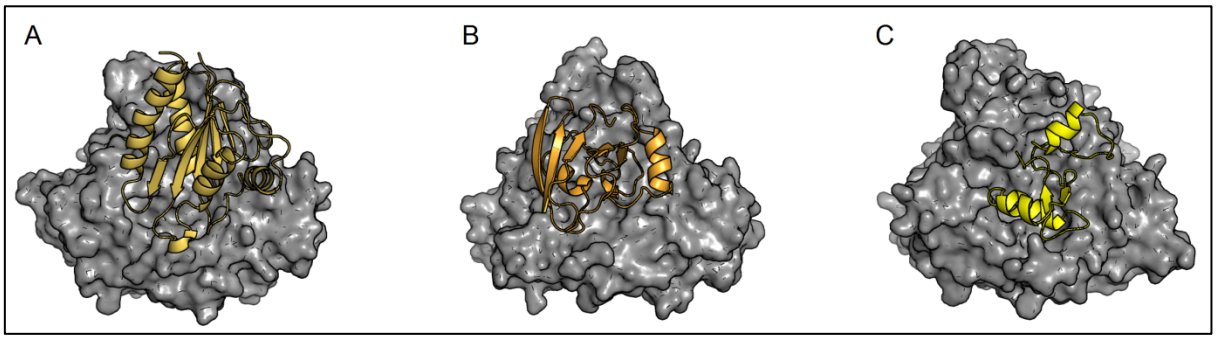


Figure 5: CYP-redox partner crystal complexes, the redox partners interact with a similar part of CYPs proximal surface. CYP102A1 heme and FMN domains (A, PDB: 1BVY), CYP101A1 with full length putidaredoxin (B, PDB: 3W9C) and CYP11A1 with partial adrenodoxin (C, PDB: 3N9Y).

1.2 Protein engineering

Protein engineering is a process aiming at improving selected properties of a protein by altering its sequence and structure.[47] Protein engineering can be divided into 3 steps: selection of the protein sites to change, introduction of the changes and screening of variants.[48] Here the first step – selection of the best sites for changes will be introduced. This part of the protein engineering process can benefit from computational analyses, like the ones presented in this dissertation. One can organize protein engineering strategies into two major groups: **random mutagenesis** and **rational design**. As the names suggest, random mutagenesis is based on introduction of amino acid mutations at arbitrary sites of the protein, whereas

rational design points to specific regions or positions of the protein with a hypothesis proposing their significance for improvement of a specific property.

The most prominent protein engineering strategy using random mutagenesis is **directed evolution**. The method consists of two iteratively repeated steps, random mutagenesis and screening for promising variants.[49,50] Since the total library size is directly related to the success of the method, it requires efficient high-throughput screening assays. Nevertheless with limited number of directed evolution cycles the number of variants necessary for screening is not prohibitive and the method proved to be successful on many occasions.[51]

Insights gathered from directed evolution studies, functional studies of enzymes, crystal structures and systematic analyses of protein families, provide constant expansion of knowledge about the biocatalysts and carry a promise of highly effective rational methods requiring much less screening effort.[48,52,53] **Rational protein design** can be divided into structure and sequence-based. In general rational design is supported by computational analyses of the of interest protein and its homologues, and aims at understanding and improving its properties.[54] **Structure-based rational design** efforts are built on the studies of protein crystal structures or homology models and their interactions with ligands or other proteins.[55] Based on this strategy it is possible to identify specific regions and positions of the protein that are interacting with the substrate and therefore influence the enzymes selectivity, specificity or activity.[56] Structure based approaches can be also used for protein stabilization. Molecular dynamics simulations can aid protein stabilization efforts by allowing for identification of flexible regions in the proteins or virtual screening of protein variants to find end eliminate destabilizing mutations.[57] The methods behind the structure-based protein design, like molecular dynamics simulations and molecular docking are also often used to understand molecular basis of selectivity and specificity of enzymes. In the chapter 3.6 a method that can be used to estimate selectivity of CYPs is applied. The method can be used to elucidate the molecular basis of selectivity, and for virtual screening of CYP libraries. **Sequence-based rational design** methods are based on comparison of amino acid sequences between homologous proteins and identification of functionally relevant positions.[53,58,59] Sequence-based methods have relatively low computational requirements and are often used to provide additional data to other design methods. Sequence-based methods were used in this dissertation to improve the CYP-redox partner interactions (chapters 3.3 and 5.3), re-design

linkers for fusion protein (chapter 3.4), and design thermostable variants of adrenodoxin reductase[b] (chapter 3.5).

The rational protein design methods are often combined with site-saturation mutagenesis[c] in so called **"semi-rational protein design"** efforts.[60,61] Presented in chapter 3.6 library of CYP101A1 variants[d] was generated under the premises of semi-rational design. Based on structure of the enzyme and known from the literature selectivity determining positions, set of residues was selected and subjected to the site-saturation mutagenesis. The semi-rational protein design combines benefits of both rational design and random mutagenesis: it requires limited screening efforts in comparison to directed evolution, but does not require hardcore computational analyses to pin-point specific interaction determining properties of interest. Since semi-rational design generates protein libraries, it is especially suitable to projects where panels of substrates are of interest.

Activity, selectivity and stability are major areas of interest in protein engineering of CYPs. Limitations in those fields have been described as crucial to overcome for CYPs to be successfully applied in industrial processes.[11,62,63] All of the protein engineering technics described in this chapter have been successfully applied to improve activity,[60,64] selectivity,[30,65] specificity[58,66] and stability[67,68] of CYPs. Nevertheless, these efforts are dispersed in the abundance of studies on CYPs, and while numerous reviews on those topics allow to keep track of the advancements in the field,[10,11,28,62,69,70] it is often difficult to transfer this knowledge to the enzyme of interest. Therefore, by systematically analyzing sequence, structure and literature information about CYPs, this dissertation aims at providing tools and guidelines to aid rational protein engineering of CYPs.

[b] Adrenodoxin reductase (AdR) together with adrenodoxin are redox partner of class I mitochondrial CYPs
[c] Site-saturation mutagenesis is a method allowing for substitution of a specific protein site with all 20 amino acids at once.[262]
[d] Generated by Dr. Paul Kelly at the University of Manchester

2. The aim of this work

The aim of this work was to establish strategies for protein engineering of CYPs. The strategies comprising this dissertation aim at overcoming shortcomings in stability, selectivity and activity of the cytochrome P450 monooxygenase systems. This aim was achieved thanks to conclusions drawn from the systematic sequence, structure and literature analyses, as well as molecular modeling. The work was done in the framework of following projects:

- establishing of class-specific standard numbering schemes for cytochrome P450 monooxygenases, and amino acid conservation analysis (chapters 3.1 and 5.1),
- identification of universal selectivity determining positions by systematic literature mining (chapters 3.2 and 5.2),
- identification of redox partner interaction sites in cytochrome P450 monooxygenases (chapters 3.3 and 5.3),
- establishing of a strategy for re-designing of the linker region in cytochrome P450 fusion systems (chapter 3.4),
- systematic approach to sequence-based thermostabilization of bovine adrenodoxin reductase (chapter 3.5),
- molecular modeling of CYP101A1 stereoselectivity towards methylated ethylbenzene-derivatives (chapter 3.6).

3. Results

The results presented here focus on the new insights into the cytochrome P450 monooxygenase (CYP) systems, which allowed establishing strategies for protein engineering of CYPs. Those strategies are based on the research enabled by application of already existing and new methods to analyze sequence, structure and function of this enzyme family. Parts of this thesis were already published or submitted for publication, and can be found in the chapter 5.

3.1 Class-specific numbering schemes for cytochrome P450 monooxygenases

In the study "Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance"[31] (complete results and discussion on this subject are included in the chapter 5.1), two class-specific numbering schemes for CYPs were established and used to perform the first conservation analysis of a large set of CYP sequences. This work was done under supervision of Prof. Dr. Jürgen Pleiss and in collaboration with Dr. Constantin Vogel[e], who is the co-author of the standard numbering scheme generation methodology.[71] Dr. Vogel significantly contributed to the generation of the numbering schemes for CYPs by providing technical expertise in the database handing and especially the numbering scheme methodology. The standard numbering schemes for protein families allow for unambiguous assignment of the standard positions to a group of proteins (Figure 6 on the following page). This in turn allows for sequence-based identification of structurally corresponding positions even in protein with low sequence identities, and enables communication about protein residues with use of family-wide standard position numbers. The numbering schemes implemented into the Cytochrome P450 Engineering Database (www.CYPED.BioCatNet.de) not only allowed for identification of the conserved positions, but also constitute pillars for the analyses presented in the subsequent chapters.

[e] University of Stuttgart

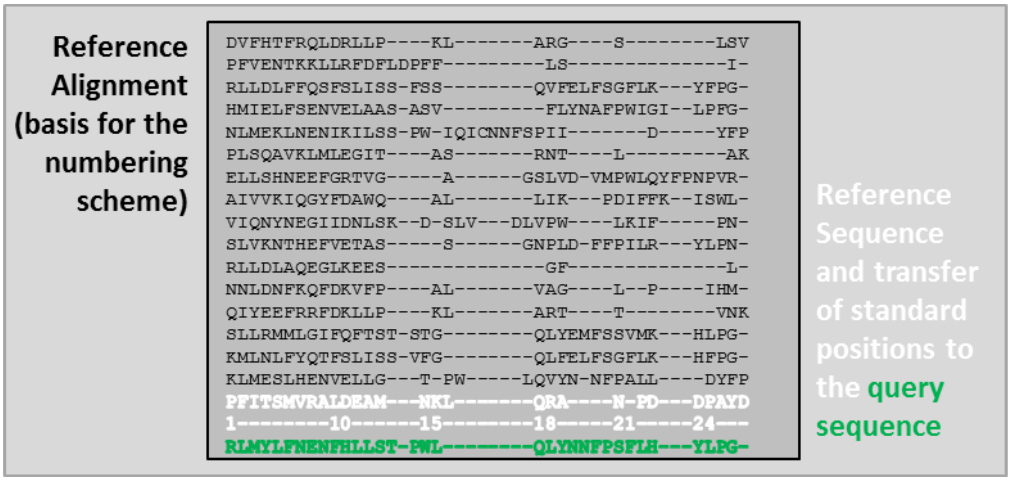


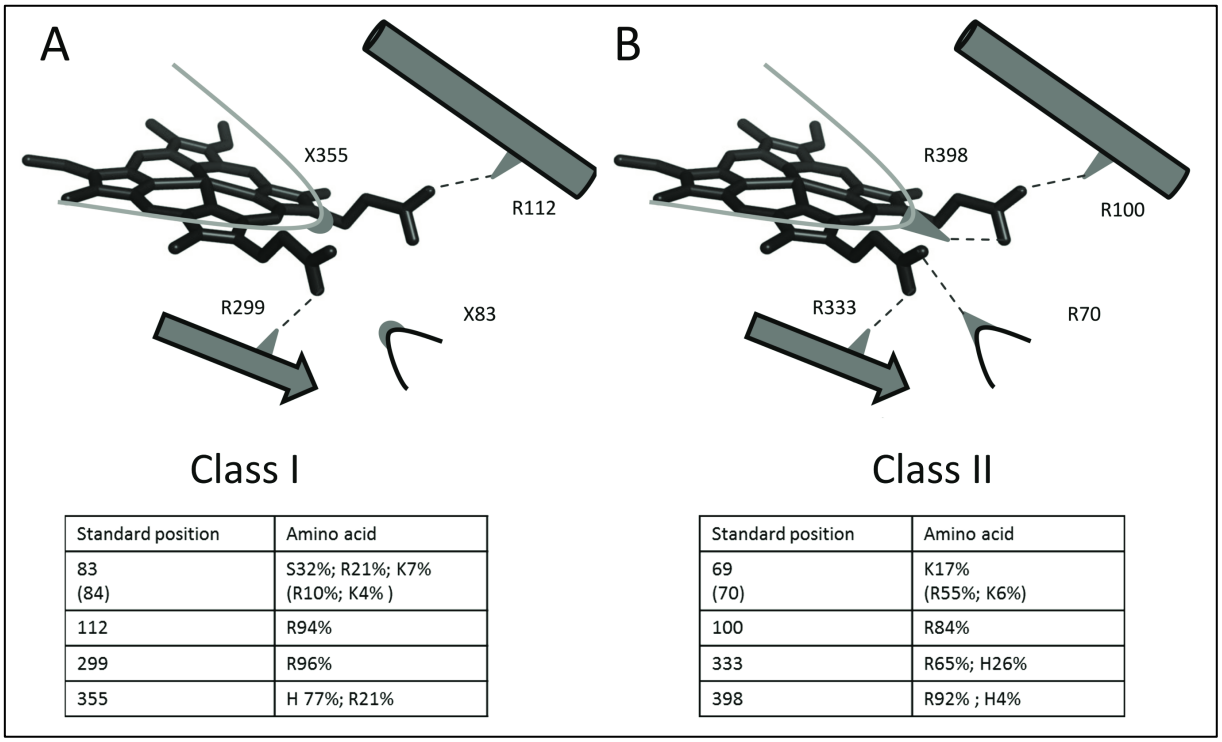
Figure 6: Functioning of a standard numbering scheme. A structure-guided reference alignment is the basis for a standard numbering scheme. The reference alignment contains a reference sequence which is the donor of standard position numbers – all positions in the protein family are numbered after this protein. To number any protein in the family, its sequence (query sequence) is aligned to the reference sequence using a hidden markov model generated from the reference alignment. The standard positions are subsequently transferred from the reference sequence to the query sequence.

Because of different structural features of class I and class II CYPs, two standard numbering schemes for CYPs were generated.[72] Differences in the length of loops and presence or absence of α-helix J' in different groups of CYPs prohibited generation of a sequence profile robust enough to cover diversity of the whole family. The most studied representatives of CYPs were selected as the reference sequences: CYP101A1 from *Pseudomonas putida* (P450cam) and the heme domain of CYP102A1 from *Bacillus megaterium* (P450BM-3), for class I and class II, respectively. To allow for conversion between class I and class II numbering schemes, a conversion table based on a structure-guided sequence alignment of the reference proteins was established (Table S1 on pages 143-145).

To present utility of the standard numbering schemes, amino acid conservation analysis of CYPs was performed. The analysis covered a set of sequences collected in the updated version of the CYPED. The database included 16732 sequences representing 13478 proteins and 408 PDB structures representing 72 proteins. All sequences were assigned into classes according to the simplified version of family classification by Hannemann et al. (full description in chapter 1.1.3, page 13),[26] which resulted in 3776 class I and 12113 class II sequences. The amino acids in all sequences were numbered accordingly to the class-specific standard numbering schemes, and amino acid frequency was calculated for each standard position. A position was considered conserved if a single amino acid or amino acids with

certain properties (aromatic, hydrogen-binding, charged, hydrophobic) were present in more than 80% of all sequences in the respective class. The conservation analysis revealed 16 and 17 positions with single amino acid conserved in class I and class II, respectively, 2 and 8 positions with aromatic amino acids, 3 positions with hydrogen-binding amino acids, 12 and 9 positions with charged amino acids, 30 and 20 positions with hydrophobic amino acids. The analysis of identified conserved positions showed that 15 were not described in the literature for CYP101A1 or CYP102A1. The identified conserved positions comprise a set of functionally or structurally relevant residues, which should not be mutated in the protein engineering efforts. This is due to the high risk of disturbing the proteins function and structure. However, mutations at those positions might be positive in outlier sequences where the conserved amino acid is not present. The set of conserved positions in CYPs is basis for one of the guidelines to protein engineering of this enzyme family.

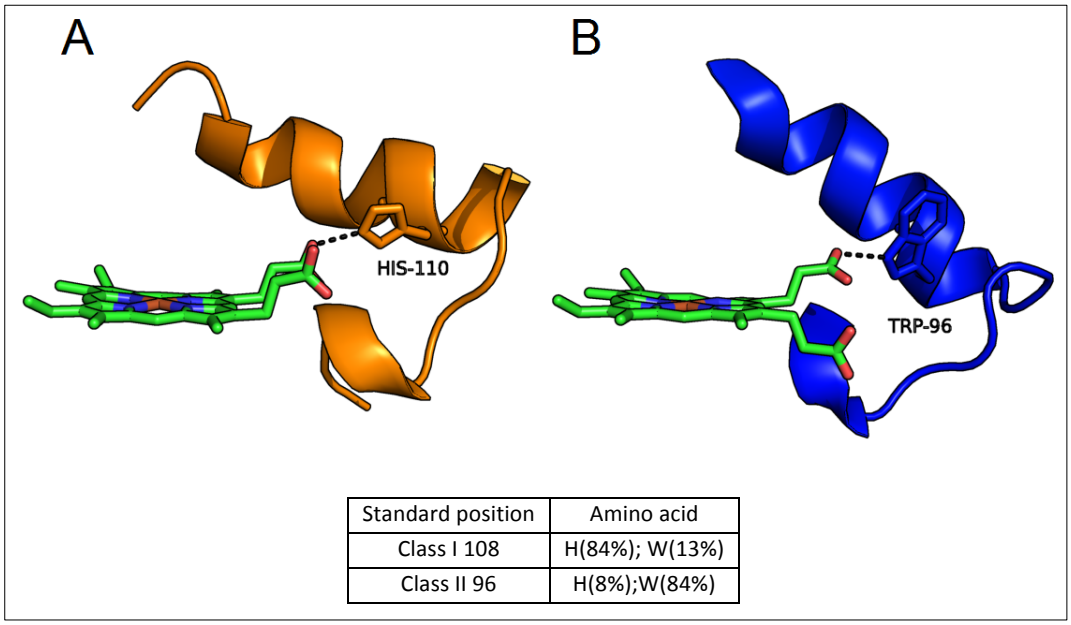
Comparison of the conservation analysis results between the two classes revealed interesting class-specific differences in the heme interacting: charged residues at standard positions 83, 112, 299, 355 and 69, 100, 333, 398 in class I and class II, respectively (Figure 7 on the following page), and histidine/tryptophan at positions 110 and 96 in class I and class II, respectively (Figure 8 on the following page). Class-specific conserved charged amino acids were described to stabilize the heme, influence its redox potential or to be part of the electron transfer route.[18,38,73–75] The differently conserved histidine/tryptophan residue was described in CYP102A1 to stabilize the heme and be involved in its incorporation.[76] Differences in amino acid conservation at those standard positions suggest functional differences in those areas in the two classes. Those differences may be important for mediating the heme redox potential and electron transfer, to be compatible with different redox partners. The conservation analysis also revealed few conserved positions on the CYP proximal surface, but with the knowledge from the further analyses (chapter 5.3) these residues are not regarded to be relevant for the CYP-redox partner interactions.



Standard position	Amino acid
83 (84)	S32%; R21%; K7% (R10%; K4%)
112	R94%
299	R96%
355	H 77%; R21%

Standard position	Amino acid
69 (70)	K17% (R55%; K6%)
100	R84%
333	R65%; H26%
398	R92% ; H4%

Figure 7: Schematic representation of conserved arginine residues in the heme environment in A) class I and B) class II CYPs. Heme interacting residues corresponding between the classes are in the same rows of the tables. In class II CYPs, heme propionate A can be involved in interaction with lysine at position 69 or with arginine at position 70. The amino acid distribution at position 70 in class II and the corresponding position 84 in class I is shown in parenthesis.



Standard position	Amino acid
Class I 108	H(84%); W(13%)
Class II 96	H(8%);W(84%)

Figure 8: Heme interacting class specific histidine or tryptophan in class I (A) and class II (B) respectively. Class I is represented by the structure of CYP108A (P450terp, PDB: 1CPT – position 110 is corresponding to standard position 108) and class II by the structure of CYP102A1 (P450BM-3, PDB: 1BVY).

3.2 Universal selectivity-determining positions in cytochrome P450 monooxygenases

In the study "Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining"[77] (complete results and discussion on this subject are included in the chapter 5.2), a sequence-based literature mining algorithm (SBLMA) was developed and used to find the most frequently mentioned in the literature structurally corresponding positions of CYPs. The identified positions are located on substrate recognition sites and influence selectivity among a diverse set of CYPs. This work was done under supervision of Prof. Dr. Jürgen Pleiss, and in collaboration with Dr. Constantin Vogel[f], who implemented storing of literature mining results into the CYPED data model[78] and displaying them on the web interface. The sequence-based literature mining algorithm is tightly implemented into the CYPED and uses its sequence data as a basis for the literature mining. **The algorithm can be divided into four steps (Fig 20 on page 100).** In **the first step**, the algorithm uses the descriptions of sequences from the CYPED to extract the CYP names[g]. In **the second step**, the PubMed database[79] is searched to find abstracts mentioning the CYP names from the CYPED. Subsequently, the abstracts mentioning the CYP names and if possible corresponding supplementary materials and full text articles are acquired. In **the third step** of the algorithm, all acquired literature files pass through custom filters identifying mentions of amino acid positions and mutations (e.g. F87A, F87, phenylalanine 87, Phe87, Phe at position 87). In **the fourth step**, the identified positions are matched to the sequence corresponding to the CYP name used to acquire the article from which they originated. The positions are rejected if the amino acid mentioned in the literature at a certain position is not matching to the one on the respective sequence. Afterwards the list of positions found in the literature is manually verified, to check if the publication is in fact mentioning the respective CYP. The complete list of sequences, CYP names, positions found in the literature, and PubMed identifiers is parsed to the CYPED. Standard numbering schemes implemented into CYPED were used for analysis of the positions mentioned in the literature. This allowed for reliable comparison of all sequences to find the structurally corresponding positions, and facilitated communication about the most frequently mentioned positions. Class I standard potions were converted to class II standard positions using the previously described conversion table (Table S1 on pages 143-145).[31]

[f] University of Stuttgart
[g] According to the Nelsons' nomenclature (i.e. CYP101A1) and based on the P450 prefix (i.e. P450cam)

Before the literature mining was performed, the cytochrome P450 engineering database was updated to contain currently available sequences and structures of CYPs. The updated resulted in 52674 sequences corresponding to 41513 sequences and 595 PDB structures corresponding to 119 proteins. This was more than three-fold increase in the number of sequences and over 60% increase in the number of proteins with known structure, since the last update (two years before). The SBLMA allowed to find over 53000 scientific articles mentioning CYPs collected in the CYPED, among which 2400 articles contained mentions of 4000 residues of 168 CYPs. The found residues could be assigned to 440 structurally corresponding standard positions of the CYP fold, which covers 96% of all standard positions.

To test completeness of the gathered data, the results of SBLMA for CYP102A1 (P450BM-3) were compared to an extensive list of CYP102A1 mutations published in a review from mid of 2011.[80] The SBLMA found 198 CYP102A1 positions to be mentioned in the literature, whereas the review cited 179 mutations. The algorithm identified 82 positions, which were not found in the review, whereas the review cites 63 positions not found with the SBLMA. Those results suggest that there are at least 261 CYP102A1 positions mentioned in the literature. The SBLMA failed to find 24% of positions mentioned for the benchmark CYP, which is an estimate for the general completeness of the gathered data. The false negatives were result of the restricted access to full text articles. It was possible to download full text articles only 22% of all identified articles.

Almost 4000 residues from 168 CYPs mentioned in the literature are represented by 440 structurally corresponding standard positions of the CYP-fold. Less than 10% of those positions were located on loops which could not be assigned to the standard positions based on the CYP102A1 numbering. The most frequently mentioned positions were found to be part of the substrate recognition sites (SRSs) and the cysteine pocket (Figure 9 on page 30). The average number of CYPs mentioned per standard position for the 98 SRS positions was 18, for the 14 cysteine pocket positions it was 14.5 whereas for the 407 remaining positions it was only 5. In the substrate recognition sites there were 38 positions mentioned for more than 18 CYPs: thirteen positions from the SRS1, five SRS2 positions, one SRS3 position, nine SRS4 positions including four conserved position, seven SRS5 positions including one conserved position and three SRS6 positions. Two conserved cysteine pocket positions (out of 14) and only four positions (out of 328) outside SRSs and the cysteine pocket were mentioned in the literature for more than 18 CYPs. This ranking shows high significance of SRS1, 4 and 5

which encircle access to the heme. I have proposed that the SRS positions, which were mentioned in context of the highest number of CYPs, most probably, also influence selectivity of a wide range of CYPs and can be announced “universal selectivity-determining positions”. The influence of those positions on selectivity of CYPs with low sequence identity was confirmed by manual validation of the found articles (chapter 5.2). The universal selectivity determining positions constitute one of the guidelines for protein engineering of CYPs. Utilization of those positions can be especially helpful in protein engineering efforts of CYPs without known structure, for which it is not possible to study interactions between the enzyme and substrate of interest. The set the universal selectivity determining positions can be identified in any CYP with use of the standard numbering scheme tool available via the CYPED web interface (www.CYPED.BioCatNet.de).

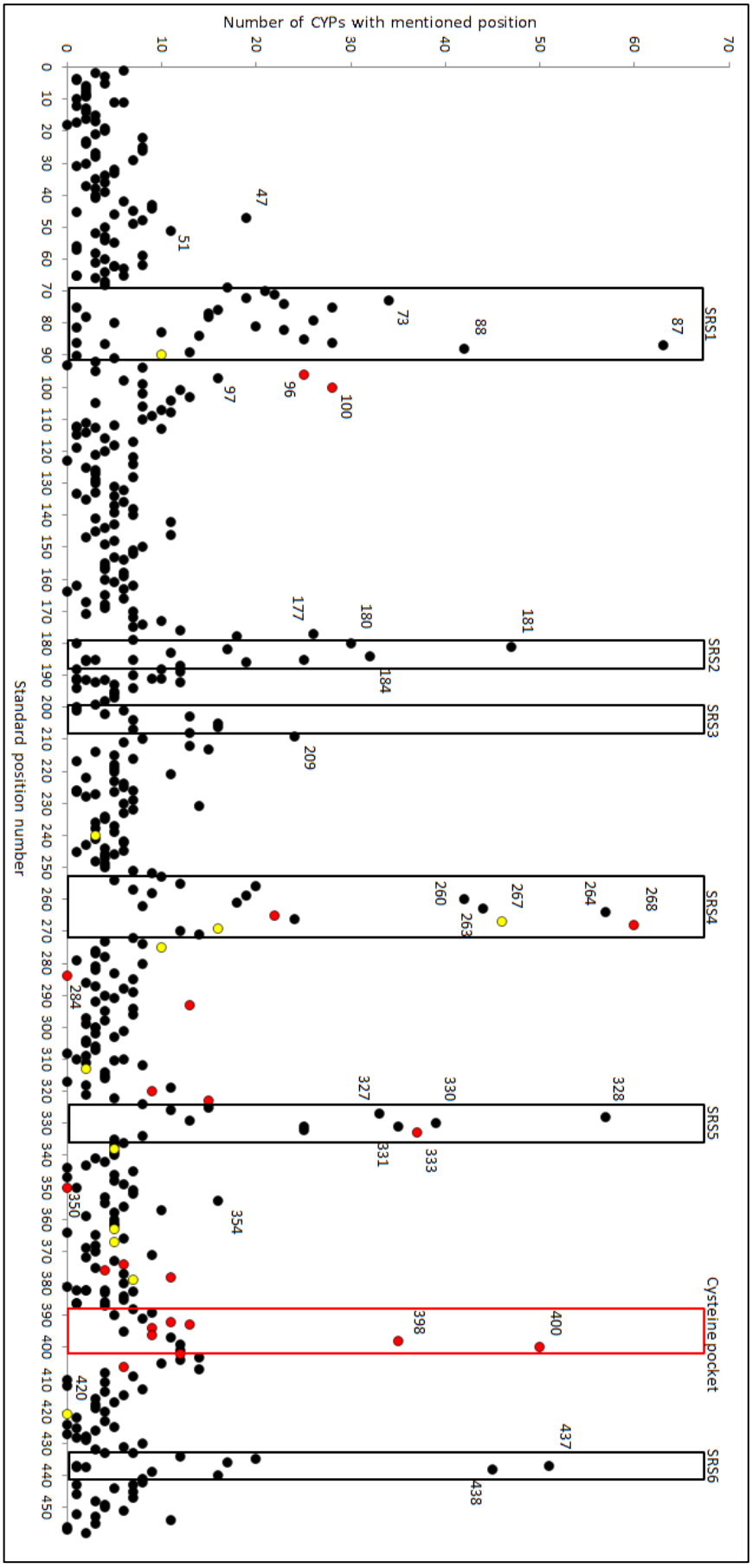


Figure 9: Plot showing the number of CYPs with described standard positions in articles about polymorphism. The substrate recognition sites and the cysteine pocket are covered by boxes. Conserved positions were colored, the positions with single conserved amino acids are red, and positions where certain properties (aromatic, charged or hydrogen binding) are conserved are colored yellow.

3.3 Redox partner interaction sites in cytochrome P450 monooxygenases

In the study "Redox partner interaction sites in cytochrome P450 monooxygenases: *in silico* analysis and experimental validation"[81] (complete results and discussion on this subject are included in the chapter 5.3), redox partner interaction sites in CYPs were identified and a generic strategy for improvement of the electron coupling efficiency and catalytic activity based on modifications of the CYP-redox partner interactions was introduced. This work was done under supervision of Prof. Dr. Jürgen Pleiss, and in collaboration with Dr. Martin J. Weissenborn[h], Sara M. Hoffmann[h], Niels Borlinghaus[h] and Prof. Dr. Bernhard Hauer[h] who performed the experimental characterization of the designed variants. This project was built on the fundaments set by the work presented in the two previous chapters. The standard numbering schemes allowed for identification of structurally corresponding positions between diverse CYPs, and the literature mining provided a list of proximal surface positions mentioned in the literature. Regions involved in the CYP-redox partner interactions were identified based on the analysis of the literature information and called redox partner interaction sites (RPISs). The strategy for improvement of the CYP-redox partner interactions takes advantage of the identified RPISs and numbering schemes, which allow for transferring of the structurally corresponding hot spot positions between all CYPs.

Described in the literature interfaces between mammalian class II CYPs and the cytochrome P450 reductases (CPR) were analyzed to identify the RPISs. Mammalian class II CYPs were selected as a basis for the analysis because of their high sequence diversity and the fact that multiple class II CYPs from one organism accept electrons from a single reductase. Because of their high similarity to human CYPs there is a large amount of published information about those enzymes. The SBLMA results were analyzed to find positions of the CYPs proximal surface, which were described in the literature to influence the mammalian CYP-CPR interactions (Table 1 on page 33). The literature mining revealed 25 class II standard positions corresponding to 45 residues described to influence the CYP-CPR interactions in 10 mammalian CYPs.

To check if the described positions are conserved in the set of 47 human class II CYPs accepting electrons from a single human CPR, the amino acid conservation analysis of those positions was performed. The analysis showed no significant conservation besides the known

[h] University of Stuttgart

partially surface accessible, heme interacting residues at standard position 100 identified in the conservation analysis of all CYPs,[31] and the hydrophobic residue at standard position 112. Similarly, the analysis of the whole proximal surface did not reveal more conserved positions specific to this group of CYPs. This surprisingly shows that the CYP-CPR interactions are not conserved in the set of human CYPs accepting electrons from a single reductase. The conservation of the human RPIS positions was compared to the conservation of all class I and class II CYPs (Table 1 on the following page). The results revealed that the average difference in amino acid frequency at those positions between 47 human class II CYPs and 11000 class II CYPs was 10% (excluding the outlier positions 386 and 386.1). The RPIS residues are similarly conserved in all class II CYPs suggesting similar significance as in human class II CYPs. The amino acid frequencies at those positions in class I differed on average 21% from human class II CYPs. Those results were expected because of differences in fold and size of class I (ferredoxins) and class II (diflavin reductases) redox partners interacting with CYPs. This difference in amino acid composition suggests some divergence in the CYP-redox partner interaction sites.

Table 1: The set of mammalian class II positions described to be involved in the CYP-CPR interactions. The table consists of class II standard numbering positions, corresponding CYP names and amino acid positions, amino acid property at the standard position, conservation of the property among human class II, class II and class I CYPs and redox partner interaction site number. Amino acid properties were assigned as follows, positive charge: H, K, R, negative charge: D, E, polar: C, N, Q, S, T, hydrophobic A, G, I, M, L, P, V.

Class II std. pos. number	CYP name and amino acid position	Amino acid property	Conservation of the property in human CYPs	Conservation of the property in class II CYPs	Conservation of the property in class I CYPs	Redox partner interaction site
59	CYP1A1(K94)[82], CYP3A4(K91)[83]	positive charge	82%	60%	43%	1
63	CYP1A1(K99)[82],CYP17A1(K89)[84], CYP19A1(K108)[85]	positive charge	24%	32%	10%	1
65/65.1	CYP2D6(E96)[86],CYP3A4(Y99)[83], CYP3A4(C98)[87]	negative charge /aromatic/polar	43%/6%/28%	30%/6%/32%	20%/1%/13%	1
68	CYP1A1(K105)[82]	positive charge	16%	9%	7%	1
97	CYP2B4(R122)[88], CYP3A4(K127)[83]	positive charge	59%	67%	33%	2
99	CYP1A1(R135)[82]	positive charge	35%	22%	21%	2
100	CYP1A1(R136)[82], CYP3A4(R130)[83], CYP2B1(R125)[89], CYP1A2(R137)[90]	positive charge	92%	94%	98%	2
101	CYP1A1(R137)[82], CYP2B4(R126)[88], CYP2C9(R125)[91]	positive charge	69%	67%	54%	2
104	CYP3A4(S134)[83]	polar	37%	45%	57%	2
108	CYP2B4(R133)[88], CYP2D6(R140)[86]	positive charge	59%	29%	17%	2
110	CYP2B4(F135)[88], CYP21A1(R132)[92]	aromatic/ positive charge	14%/10%	13%/16%	1%/61%	2
112	CYP2B4(M137)[88]	hydrophobic	90%	74%	91%	2
113	CYP2B4(K139)[88], CYP3A4(K143)[83]	positive charge	57%	50%	30%	2
236	CYP2B4(V267)[93]	hydrophobic	55%	53%	50%	3
239	CYP2B4(V270)[93]	hydrophobic	51%	23%	27%	3
305	CYP3A4(Y347)[83]	aromatic	47%	38%	GAP	4
310	CYP17A1(R347)[84]	positive charge	43%	26%	GAP	4
319	CYP17A1(R358)[84]	positive charge	73%	54%	17%	4
383	CYP2B4(L420)[93], CYP19A1(K420)[85]	hydrophobic/ positive charge	8%/16%	30%/13%	GAP	5
386/386.1	CYP1A1(K440)[82], CYP2B4(R422)[88],	positive charge	55%/63%	17%/5%	GAP	5
388	CYP3A4(Y430)[83]	aromatic	24%	40%	GAP	5
397	CYP1A1(K453)[82], CYP2B4(K433)[88], CYP2D6(R440)[86]	positive charge	51%	45%	7%	6
399	CYP1A1(R455)[82]	positive charge	12%	10%	29%	6
404	CYP3A4(R446)[83]	positive charge	31%	26%	27%	6
407	CYP1A1(K463)[82], CYP2B4(R443)[88], CYP19A1(R449)	positive charge	41%	24%	62%	6

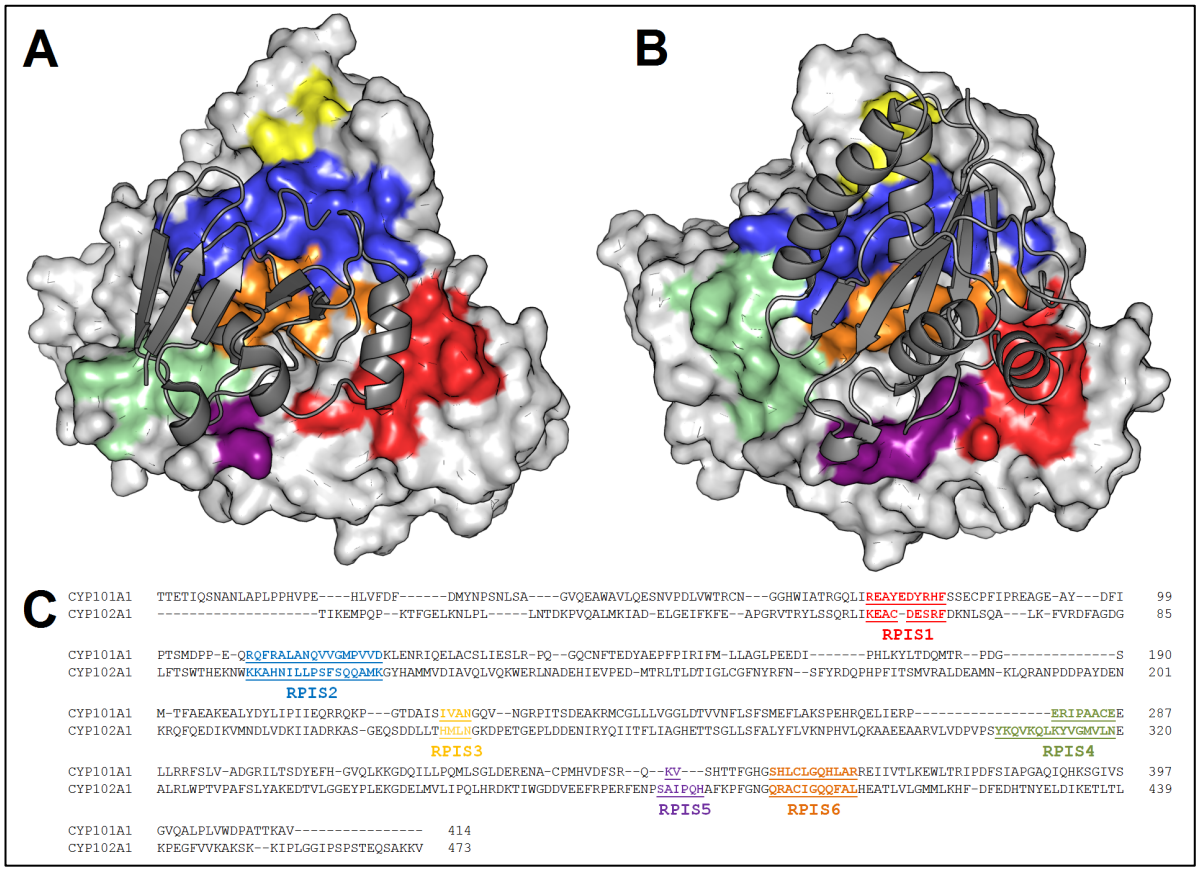


Figure 10: Reductase interaction sites on the proximal surfaces and sequences of CYP101A1 (PDB code: 3W9C[45]) and CYP102A1 (PDB code:1BVY[43]). The RPISs are highlighted in different colors: RPIS1 red, RPIS2 blue, RPIS3 yellow, RPIS4 green, RPIS5 purple, and RPIS6 orange. A: CYP101A1 structure with co-crystalized putidaredoxin (showed as grey cartoon), which rests over most of the RPISs. B: CYP102A1 structure with co-crystalized FMN domain of the reductase (showed as grey cartoon), which rests right over the RPISs. C: structure-guided sequence alignment by STAMP[94] between CYP102A1 (PDB code: 1ZOA)[95] and CYP101A1 (PDB code: 1PHG)[96] with marked RPISs.

The positions identified to influence CYP-redox partner interactions in mammalian class II CYPs are located on six RPIS occupying most of the CYPs proximal surface, where putidaredoxin binds to the CYP101A1 and where the FMN domain of the CYP102A1 reductase binds to the heme domain (Figure 10). The structures confirm that regions identified for human class II CYPs are also relevant for the CYP-redox partner interactions in other enzymes. Additional evidence based on literature data confirming the relevance of RPISs for diverse CYPs and redox partners are presented in the chapter 5.3.5 (discussion section of the "Redox partner interaction sites in cytochrome P450 monooxygenases: *in silico* analysis and experimental validation" manuscript). The literature data available about modifications of CYP-redox partner interactions in class I and class II enzymes, suggests that four RPISs are involved in the CYP-redox partner interactions in prokaryotic class I CYPs, five RPISs in the mitochondrial class I CYPs and all six RPISs in class II CYPs. A previously published review

from 2003 by Hlavica et al. provided an initial generalized view on the CYP-redox partner interaction, presented here analyses extend these information by providing details about class-specific differences and identification of regions involved in the CYP-redox partner interactions.[69]

To check if the identified positions also influence the catalytic activity and electron coupling of an interclass fusion system of CYP153A6 from *Marinobacter aquaeolei* (natively class I) and reductase domain of CYP102A1 from *Bacillus megaterium* (class II),[30] I have designed a generic strategy aiming at mimicking the charges of RPIS residues on the reductases natural redox partner. Charge was chosen as a property to mimic, because of its reported high significance in the CYP-redox partner interactions.[82,97,98] This strategy was designed to improve the CYP-redox partner interactions with a small number of variants. The positions for mutagenesis were selected from a set of the RPIS positions described for mammalian CYPs (Table 1 on page 33). Structurally corresponding amino acids at the RPIS positions were compared between CYP153A6 and CYP102A1 (Table 2 on the following page). The comparison was enabled by the standard numbering schemes. CYP102A1 positions (which are the class II standard positions) were compared to the CYP153A6 positions numbered using class I numbering scheme and converted to class II standard positions. Six structurally corresponding positions with different charge in the two proteins were selected for mutagenesis (Table 2 on the following page). In RPIS1, two variants were designed (L115K and S120D). Variants S122D and D125S were omitted because of similar change to the S120D and the greater distance to the center of the proximal surface (position of the proximal cysteine). In RPIS2, D153K and Q166K were designed. RPIS3 was described to be important for hydrophobic interactions and the proposed variants did not introduce hydrophobic residues. In RPIS4 and RPIS5, no positions were selected for mutagenesis because those regions are shorter in class I than class II CYPs (Figure 10 on the previous page) and reliable comparison of those regions between CYP153A6 and CYP102A1 was not possible. In RPIS6, variants R422Q and E425L were designed. Variant R399R was omitted because it is an immediate neighbor of the proximal cysteine and might disturb the enzymes' activity.

Table 2: Comparison of amino acid residues at the CYP-CPR interaction interface between CYP102A1 and CYP153A6. Table consists of class II standard position numbers, amino acids at the corresponding positions in CYP102A1 and CYP153A6, position numbers in CYP153A6 and redox partner interaction site number. Positions selected for mutagenesis are in bold and highlighted grey, CYP153A6 amino acids were replaced by CYP102A1 amino acids.

Class II std. pos. number	CYP102A1	CYP153A6	CYP153A6 pos. number	Redox partner interaction site
59	**K**	**L**	**115**	**1**
63	**D**	**S**	**120**	**1**
65	S	D	122	1
68	D	S	125	1
97	**K**	**D**	**153**	**2**
99	A	Q	155	2
100	H	R	156	2
101	N	S	157	2
104	L	Q	160	2
108	S	A	164	2
110	**Q**	**K**	**166**	**2**
112	M	L	168	2
113	K	K	169	2
236	H	L	280	3
239	N	S	283	3
305	Y	-	GAP	4
310	Q	-	GAP	4
319	L	S	345	4
383	S	-	GAP	5
386	P	-	GAP	5
388	H	-	GAP	5
397	Q	V	415	6
399	A	R	417	6
404	**Q**	**R**	**422**	**6**
407	**L**	**E**	**425**	**6**

The designed variants were subsequently generated and characterized in the lab by the experimental collaborators in this project. It was possible to express all variants except the R422Q. The experimental validation of the designed variants revealed that in all cases the activity and electron coupling efficiency were changed in comparison to the wild type (WT). The electron coupling efficiency (Table 3 on the following page) revealed that WT with 68% was not the best and that variants K166Q and D153K had improved coupling efficiency: 76% and 89%, respectively. The variants S120D and E425L did not influence coupling efficiency in a noticeable way, whereas variant L155K showed decrease in coupling. The analysis of

initial reaction rates and conversion after 1h showed that WT was more active than all RPIS variants, with the best variant K166Q having 28% lower conversion after 1h and S120D 36% lower conversion, whereas the other three variants showed less than 50% conversion as compared to the WT (Table 3). The experimental characterization of designed variants confirms influence of the identified RPIS position on activity and electron coupling in an interclass CYP chimera.

Table 3: Initial reaction rates, conversion after 1h and coupling efficiency of wild type CYP153A6-CPR and RPIS variants. The best values for each property are in bold. The measurements were performed in triplets and averaged, and standard deviations were calculated.

	WT	L115K	S120D	D153K	K166Q	E425L
Initial rate [µmol/min*µmol]	**23.2±0.7**	1.9±0.1	5.3±0.7	2.2±0.2	10.6±0.6	2.2±0.6
Conversion after 1 h [%]	**80.1***	27.4±1.7	44.4±6.0	27.2±0.4	52.8±3.7	29.8±6.1
Coupling efficiency [%]	67.8±11.5	56±1.4	72.6±4.5	76.2±5.4	**89.3±6.9**	63.1±2.8

*one sample analyzed

Identification of the redox partner interaction sites in CYPs, and establishment of a generic strategy for improvement of the CYP-redox partner interactions constitute one of the strategies for the protein engineering of CYPs. Engineering of this interaction is increasingly important because redox partners of many novel CYPs are not known, and those enzymes often are being combined with non-native redox partners to obtain catalytically active systems. Therefore, understanding of the CYP-redox partner interactions is the basis of a successful protein engineering strategy.

3.4 The impact of linker length on cytochrome P450 monooxygenase fusion constructs

In this chapter results of cytochrome P450 monooxygenase fusion protein linker re-design are presented. This work was done under supervision of Prof. Dr. Jürgen Pleiss, and in collaboration with Sara M. Hoffmann[i], Dr. Martin J. Weissenborn[i], Sandra Notonier[i], Dr. Bettina M. Nestl[i] and Prof. Dr. Bernhard Hauer[i], who initiated the project, contributed to design of the linkers and performed experimental characterization of the designed variants. CYP153A6 from *Marinobacter aquaeolei* used in this study was previously fused to redox

[i] University of Stuttgart

partner domains of different natural fusion proteins.[99] Here, one of the fusion constructs was optimized by varying the linker region.

CYP153A6 is natively a class I enzyme accepting electrons from ferredoxin and ferredoxin reductase.[100] The artificial fusion constructs of this enzyme were created with redox partners containing ferredoxin and ferredoxin reductase (class I PFOR and RhF) as well as diflavin cytochrome P450 reductase (class II CPR).[99] A CYP153A6-CPR fusion was generated by combining CYP153A6 with redox domains of CYP102A1 from *Bacillus megaterium,* coupled by a natural linker region that was extended by a 3xGGS peptide. CYP153A6-PFOR fusion was generated by combining CYP153A6 with redox domains of CYP116B3 from *Rhodococcus ruber,* natural linker was initially used in this system. CYP153A6-RhF fusion was generated by combining CYP153A6 with redox domains of CYP116B2 from *Rhodococcus sp.* and a natural linker. Initial reaction rates of crude lysates with dodecanoic acid as a substrate were compared between the three constructs. Results showed comparable activity of CYP153A6-CPR and CYP153A6-PFOR, whereas CYP153A6-RhF exhibited noticeably lower initial reaction rate (Table 4). The CYP153A6-PFOR construct was selected for optimizations by varying the linker length because like the natural CYP153A6 redox partner it also contains a ferredoxin domain delivering, which is delivering electrons to the CYP.

Table 4: Initial reaction rates of crude extracts of CYP153A6 fusion constructs with dodecanoic acid as a substrate. The measurements were performed in triplets and averaged, and standard deviations were calculated, as described in chapter 5.3.

	CYP153A6-CPR	CYP153A6-PFOR	CYP153A6-Rhf
Initial reaction rate [% s-1]	2.3 ± 0.4	2.1 ± 0.08	0.7 ± 0.03

To learn more about the linker regions and establish basis for the rational design, previously established version of the Cytochrome P450 Engineering Database was analyzed.[31] Fusion proteins homologous to the CYP116B3 (NCBI gi: 62869557) from the homologous family CYP116B were aligned using the Clustal omega alignment program.[101] The numbering scheme-based alignment was not used in this case because the class-specific numbering schemes are limited only to the CYP-domains, and do not cover linker regions. The linker region was identified in 52 sequences of the CYP116B homologous family by annotating alignment columns corresponding to the CYP116B3 linker sequence (Figure 11A on the following page).[102]

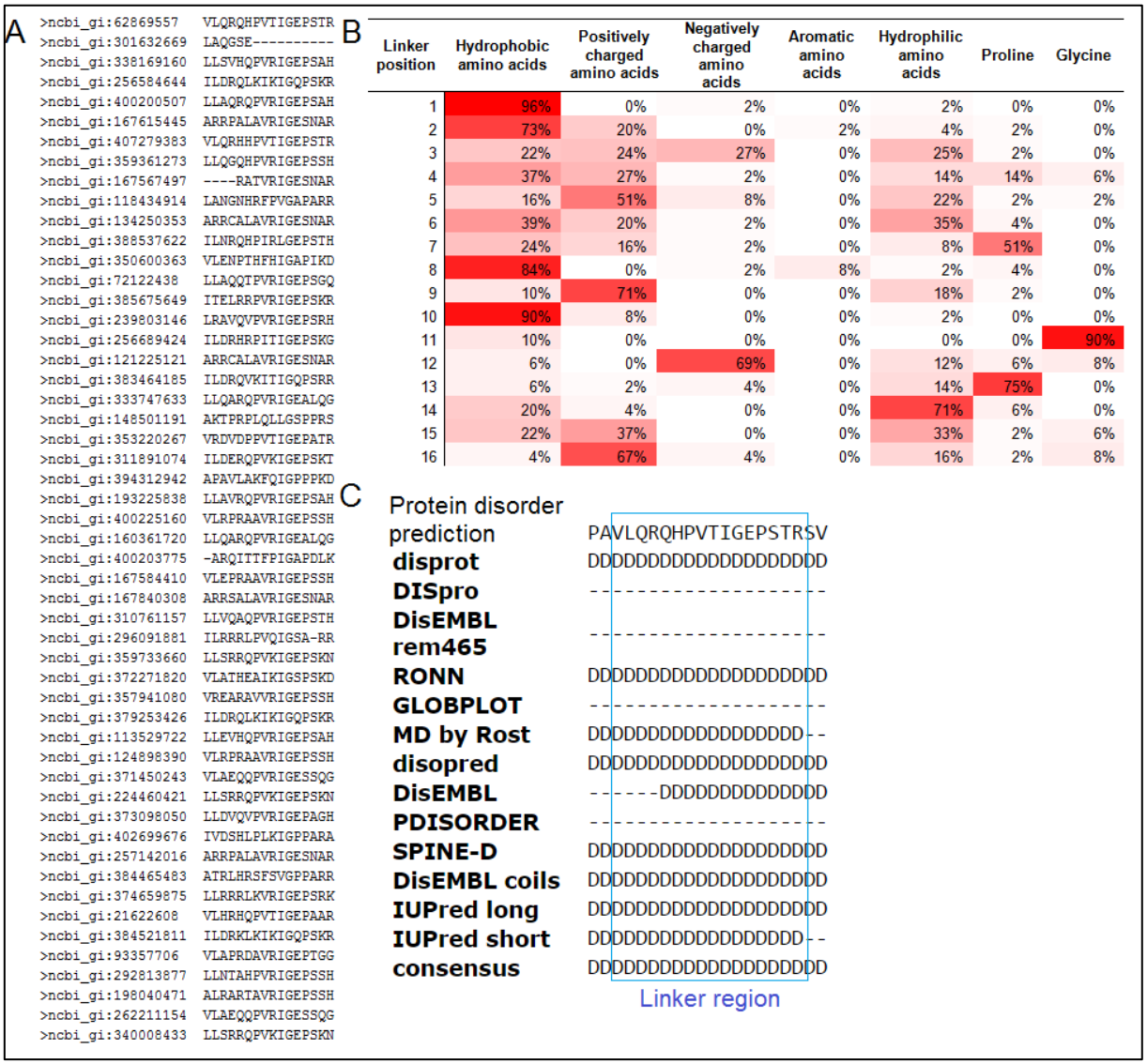


Figure 11: Sequence analyses of the CYP116B homologous family linker regions. A: Sequence alignment of CYP116B homologous family members from CYPED. Only columns corresponding to the linker regions are presented. B: Conservation of amino acid properties in the liker region of CYP116B homologous family. Amino acids were assigned to seven groups based on the physiochemical properties: hydrophobic (AILMVC), positively charged (HKR), negatively charged (DE), aromatic (FWY), hydrophilic (NQST), proline and glycine. C: Prediction of the protein disorder of CYP153A6-PFOR with natural linker performed with the GeneSilico metaserver. Consensus of the different servers is presented in the last row. The Linker region is in the blue box

Amino acid conservation analysis was performed for the linker region and conservation of amino acid properties for each linker position was also measured (Figure 11B). The results revealed that positions 1-2 and 8-16 of the linker had conserved amino acid properties (>70%) whereas region 3-7 was more variable (<50%). Protein disorder analysis performed using GeneSilico metaserver[103] showed high probability of disordered structure in the linker region, which suggests that no secondary structure elements should be disturbed by modifications of

this region (Figure 11C).[103] The sequence of CYP153A6-PFOR with its natural linker was used for the protein disorder analysis.

Based on the knowledge gained by the sequence analyses a generic strategy for linker design with redox partners from CYP116B homologous family was proposed. Since the C-terminal part of the linker region (positions 8-16) contains conserved residues and is connected to the unmodified redox partner, it was decided to leave this region unchanged. Thus, allowing to preserve possible interactions between the linker and redox partner. This part of the linker region also contains highly conserved glycine and proline residues. Those residues play an important role in the flexibility of the linker, and might be responsible for facilitating conformational changes necessary for the proper contact between the CYP and ferredoxin to occur. Additionally conserved amino acids at the linker N-termini (positions 1 and 2) were also kept unchanged. Modifications of the linker length were conducted by variation of the less conserved region. Seven linkers (L1-L7) with different length were designed, two shorter and five longer than the natural (WT) linker (Table 5).

Table 5: Different linker regions designed for the CYP153A6-PFOR fusion constructs.

Linker	Length	Sequence
L1	11	**VL**-----**VTIGEPSTR**
L2	14	**VL**QRQ--**VTIGEPSTR**
WT	16	**VL**QRQHP**VTIGEPSTR**
L3	18	**VL**VLQRQHP**VTIGEPSTR**
L4	21	**VL**QRQVLQRQHP**VTIGEPSTR**
L5	24	**VL**QRQHPVVLQRQHP**VTIGEPSTR**
L6	29	**VL**QRQVLQRQHPVVLQRQHP**VTIGEPSTR**
L7	32	**VL**QRQHPVVLQRQHPVVLQRQHP**VTIGEPSTR**

The designed variants were subsequently generated and characterized in the lab by the experimental collaborators in this project. The CYP153A6-PFOR fusion constructs with seven designed linkers (L1-L7) were genetically engineered by QuikChange© and via an oligo annealing method and expressed in BL21(DE3) *E. coli* cells. All variants were experimentally characterized and compared to the construct with a natural linker (WT) and the CYP153A6-CPR fusion protein. Initial reaction rates with dodecanoic acid as a substrate and conversion after 1h were measured *in vitro* (Table 6 on the following page). The lysates containing 0.625 µM of the CYP variants were assayed at different time points at 30°C. The

reaction was quenched with HCl and the probes were extracted and derivatized followed by GC-FID analysis.

Table 6: Initial reaction rates and conversion of dodecanoic acid after 1h measured for different CYP153A6-PFOR constructs and CYP153A6-CPR. The measurements were performed in triplets and averaged, and standard deviations were calculated, as described in chapter 5.3.

Linker construct	Initial reaction rate ($nmol_{Substrate}$/ [$nmol_{P450}$ min])	Conversion 1 h [%]
CYP153A6-PFOR L1	9.6	57.0 ± 3.6
CYP153A6-PFOR L2	8.64	92.5 ±1.9
CYP153A6-PFOR WT	6.72	88.2 ± 2.0
CYP153A6-PFOR L3	11.2	77.0 ± 1.2
CYP153A6-PFOR L4	6.08	72.9 ± 6.0
CYP153A6-PFOR L5	4.48	73.8 ± 7.2
CYP153A6-PFOR L6	5.12	79.0 ± 1.2
CYP153A6-PFOR L7	5.76	71.7 ± 1.6
CYP153A6-CPR	7.36	91.6 ± 3.7
Empty pET28 Vector	-	-

Three CYP153A6-PFOR variants (L1, L2 and L3) exhibited increased initial reaction rate in comparison to the construct with WT linker region and CYP153A6-CPR construct. Conversion after 1h was the highest and similar between CYP153A6-PFOR WT, L2 and CYP153A6-CPR. Conversion after 8h, expression, and conversion after 8h relative to the concentration of the CYP were measured *in vivo* (Table 7).

Table 7: Conversion after 8h, CYP concentrations and conversion relative to the CYP concentration of different CYP153A6-PFOR constructs and CYP153A6-CPR. The measurements were performed in triplets and averaged, and standard deviations were calculated.

Linker Construct	Conversion 8 h/ g_{cww} [%]	Concentration of CYP [mg CYP/g_{cww}]	Conversion 8h /P450 [%/mg CYP]	Reaction rate after 2h [$nmol_{Substrate}$/ ($nmol_{P450}$ min)]
CYP153A6-PFOR L1	15.13 ± 4.69	0.08	229.40 ± 71.11	22.1
CYP153A6-PFOR L2	61.24 ± 12.01	1.	67.96 ± 13.33	24.98
CYP153A6-PFOR WT	42.54 ± 6.60	0.76	66.74 ± 10.36	14.76
CYP153A6-PFOR L3	9.64 ± 2.59	0.1	109.62 ± 29.44	31.73
CYP153A6-PFOR L4	33.97 ± 11.24	0.68	59.44 ± 19.66	21.08
CYP153A6-PFOR L5	33.37 ± 18.58	0.79	50.60 ± 28.18	20.67
CYP153A6-PFOR L6	62.06 ± 12.45	1.47	50.42 ± 10.12	21.73
CYP153A6-PFOR L7	53.44 ± 17.35	1.1	57.90 ± 18.79	23.88
CYP153A6-CPR	50.54 ± 17.55	3.41	17.69 ± 6.14	9.38
Empty pET28 Vector	0	0	0	0

The expression yields varied substantially between the constructs. Variants L1 and L2 showed substantially decreased expression yield, 10 and 7-fold lower than the construct with WT

linker region, respectively. Whereas the other variants showed improvement in expression yield, CYP153A6-CPR showed the highest expression yield 3-fold higher than the CYP153A6-PFOR WT. Activity of the constructs relative to the expression yield was compared between the variants. The highest conversion per nmol CYP during the first two hours was obtained with variants L2 and L3, 1.7 and 2-fold increase in comparison to the variant with WT linker region. Even though the variant L3 showed the highest conversion per nmol CYP, its conversion after 8h was only 10% because of low expression.

The variant with linker L2 had the highest initial reaction rate, expressed with high yield, exhibited one of the highest conversions, and showed high activity *in vitro*. This variant also showed higher conversion per nmol CYP in comparison to CYP153A6-CPR and CYP153A6-PFOR WT. Variant CYP153A6-PFOR L2 has only two amino acids shorter linker than the CYP153A6-PFOR WT, which shows that in this case no dramatic changes in the linker length were necessary. The strategy for re-designing the linker region based on the natural amino acid sequence of the natural linker region and systematic sequence analysis of homologous CYPs proved successful.

The strategy presented here is part of the guidelines to protein engineering of CYPs. As mentioned in the previous chapter, redox partners of many novel CYPs are not known, and those enzymes are coupled with non-native redox partners and often combined into fusion proteins. In such fusion proteins not only the CYP-redox partner interface, but also the linker between CYP and redox partner, significantly influences CYPs activity. Therefore, the generic strategy described in this chapter is important in comprehensive protein engineering of CYPs. The method can be applied exactly as described here to fine-tune any other fusion construct with PFOR redox partner, or be easily adapted for engineering of other CYP fusion enzymes.

3.5 Thermostabilization of bovine adrenodoxin reductase by sequence consensus approach

In this chapter preliminary results of the adrenodoxin reductase thermostabilization are presented. This work was done under supervision of Prof. Dr. Jürgen Pleiss, and in

collaboration with Nina Beyer[j], Marijke Jansma[j], Dr. Hein Wijma[j] and Prof. Dr. Dick Janssen[j], who initiated the project, performed the molecular dynamics simulations and experimental characterization of the designed variants. Adrenodoxin reductase (AdR) is part of the mitochondrial class I electron transfer systems. AdR transfers two electrons from NADPH through FAD cofactor to the adrenodoxin, which one at a time shuttles them to the CYP.[26] AdR consists of FAD and NADPH binding domains, binds adrenodoxin, and requires conformational changes for its activity.[104,105] Therefore it is a challenging protein for thermostabilization, because the introduced mutations must not negatively influence any of those functions. The thermostabilization strategy used in this study was based on the sequence consensus approach.[106] The sequence consensus approach is formed on the premise that the most frequently occurring amino acid at a particular column of a multiple sequence alignment is contributing to the stability of a protein present in this alignment in a more significant way than the other amino acids.[107] The mutations proposed based on the consensus approach were also tested by short molecular dynamics simulations to check their influence on the AdR flexibility. This step was implemented as in the FRESCO approach for protein thermostabilization.[57]

The adrenodoxin reductase database was established to collect sequences homologous to the bovine AdR. Ten sequences of AdRs from different organisms (*Phytophthora infestans* gi:301115622*, Ricinus communis* gi:255573812*, Dicentrarchus labrax* gi:317419131*, Trichophyton tonsurans* gi:326469630*, Pediculus humanus corporis* gi:242018843*, Mycobacterium smegmatis* gi:399986456*, Aedes aegypti* gi:62529864*, Anopheles gambiae* gi:63148520*, Frankia sp.* gi:15667237*, Bos Taurus* gi:217434) were used as seed sequences. The database creation was performed as described before with e-value of $1e^{-50}$.[32,77] Resulting database contained 1575 sequences and 9 crystal structures grouped into 24 homologous families with sequence identity of 55%.

High quality sequence alignments are the basis for the successful application of sequence consensus approach. Therefore, four sets of sequences were used as a basis for the method, three sets with different sequence identities as compared to the bovine AdR, and one with thermostable AdR homologues. Such approach was aimed at providing insight into the relationship between the sequence identities of proteins in the alignment and effectiveness of the proposed mutations. Sequence identities between all sequences in the AdR database were

[j] University of Groningen

calculated using the Needleman-Wunsch alignment method.[108] Three sets of sequences with different global sequence identities as compared to the bovine AdR were generated: 30-60% (1137 sequences), 30%-100% (1336 sequences) and 60%-100% (199 sequences). Additionally nine sequences from organisms containing the word "thermo" in their name were also aligned to the sequence of the bovine AdR. This was aimed at checking if mutations back to thermophile consensus are on average more stabilizing than mutations proposed from the other sequence sets. Sequences from each set were aligned using the Clustal Omega alignment software.[101] Amino acid frequencies of the alignment columns corresponding to the bovine AdR sequence were calculated and amino acids found in more than 50% of the sequences were identified as the consensus. If the consensus amino acid was not identical to the bovine AdR amino acid at the corresponding position, it was proposed for the back to consensus mutation. The consensus approach for thermostabilization of bovine AdR based on the four sequence sets resulted in 98 proposed mutations (Table 8 on the following page). Two mutations (Q89P and A426G) were excluded from the further analyses, because of introduction of unfavorable amino acids into α-helices. An additional step consisting of testing by short (five repetitions of 100 ps) molecular dynamics simulations was applied to check influence of the proposed mutations on the AdR flexibility (PDB code: 1CJC[104]), 12 mutations showed significant increase in flexibility and were excluded from the experimental validation (Table 8). Additionally four variants with double mutations combining changes at neighboring positions were proposed (T6P+P7L, Q40L+L41P, Q92R+D93E, H107D+Q108R). Thus, 88 variants were designed.

Multiple proposed mutations were overlapping the four sequence sets, 57 out of 84 proposed single mutations were found in the set of sequences 30-60% identical to bovine AdR, 34 were found in the 30-100% set of sequences, 19 were found in the 60-100% set of sequences, and 25 of the proposed single mutations were found in the set of AdR homologs from thermostable organisms.

The designed variants were subsequently generated and experimentally characterized by the experimental collaborators in this project. The mutations were introduced using QuikChange© PCR, and confirmed by sequencing. Up to now, 48 variants were characterized. In this set, due to experimental errors nine variants had multiple back to consensus mutations. Thermostability of those variants and wild type AdR were tested using the ThremoFAD approach.[109]

Table 8: Mutations proposed by the sequence consensus approach for thermostabilization of bovine AdR based on four sequence sets ('1': 30-60% sequence identity with the bovine AdR, '2': 30-100%, '3': 60-100%, '4': homologs from thermostable organisms), and confirmation of no negative effects on protein flexibility by molecular dynamics ('+': no increase in flexibility, '-': increased flexibility).

AdR amino acid	AdR position	Consensus amino acid	Sequence sets	Molecular dynamics test
T	6	P	1	+
P	7	L	1	+
Q	8	R	1	+
I	9	V	1,2	+
C	10	A	1,2,4	+
V	11	I	1	+
T	21	A	1,2	+
S	30	P	3	+
A	32	V	1,2	+
Y	37	F	1	+
K	39	R	1,2	+
Q	40	L	1,2,4	+
L	41	P	1,2,4	+
F	49	Y	1	+
V	58	I	1	+
T	66	E	4	+
T	68	V	1	+
D	72	P	1,4	+
C	74	F	4	+
A	75	R	1,2	+
Q	89	P	3	-
Q	92	R	3	+
Q	92	L	4	-
D	93	E	3	+
H	96	D	1,2	-
V	99	I	1	+
S	101	A	1,2	+
H	107	D	1,2,4	+
Q	108	R	1,2,3,4	+
E	116	D	1,4	+
V	120	S	1,2	+
A	125	E	4	+
L	133	H	1,2	+
E	135	D	1,2	+
R	137	Q	3	+
R	137	A	4	-
I	150	V	1,2	+
L	151	I	1,4	+
P	167	D	4	+

AdR amino acid	AdR position	Consensus amino acid	Sequence sets	Molecular dynamics test
D	169	E	3	+
E	172	A	4	-
T	177	A	4	-
A	179	H	4	+
G	182	E	4	+
A	183	V	3	+
T	191	E	4	+
I	194	V	1	+
L	201	A	1	+
V	203	A	1,2,4	+
M	213	L	1,2,4	+
T	219	V	1	-
P	221	V	1,2	+
M	222	I	3	+
L	223	V	1	+
G	230	D	1	+
A	238	V	3	+
A	239	P	3	+
M	246	T	3	+
R	265	Q	3	+
A	271	R	1	+
Q	282	E	1,2	+
P	285	G	1,2	+
A	292	V	1,2	+
A	298	E	1	+
V	299	R	1,2,4	+
I	305	V	3	+
P	313	G	1	+
D	316	E	1	+
S	327	R	1,2	+
I	329	V	1,2,3	+
S	333	G	1,2,4	+
V	340	L	1	+
V	349	I	2,3	+
E	353	G	4	-
C	364	V	1	+
V	368	I	1,2	+
T	377	G	1,2	+
T	377	A	3	+
T	378	S	4	+
M	380	K	1,2	+
S	383	A	1,2,4	+
L	385	E	1,2	+

AdR amino acid	AdR position	Consensus amino acid	Sequence sets	Molecular dynamics test
G	387	V	1	+
H	398	L	3	-
S	401	D	4	+
S	407	Y	3	-
F	409	A	3	+
K	411	Q	3	-
D	415	L	1	-
D	415	S	3	-
W	420	R	3	+
F	424	W	1	+
D	426	G	1,2,4	+
L	430	I	1,4	+
V	435	R	4	+
A	440	G	3	-
K	443	R	1,2	+
L	449	V	2,3	+

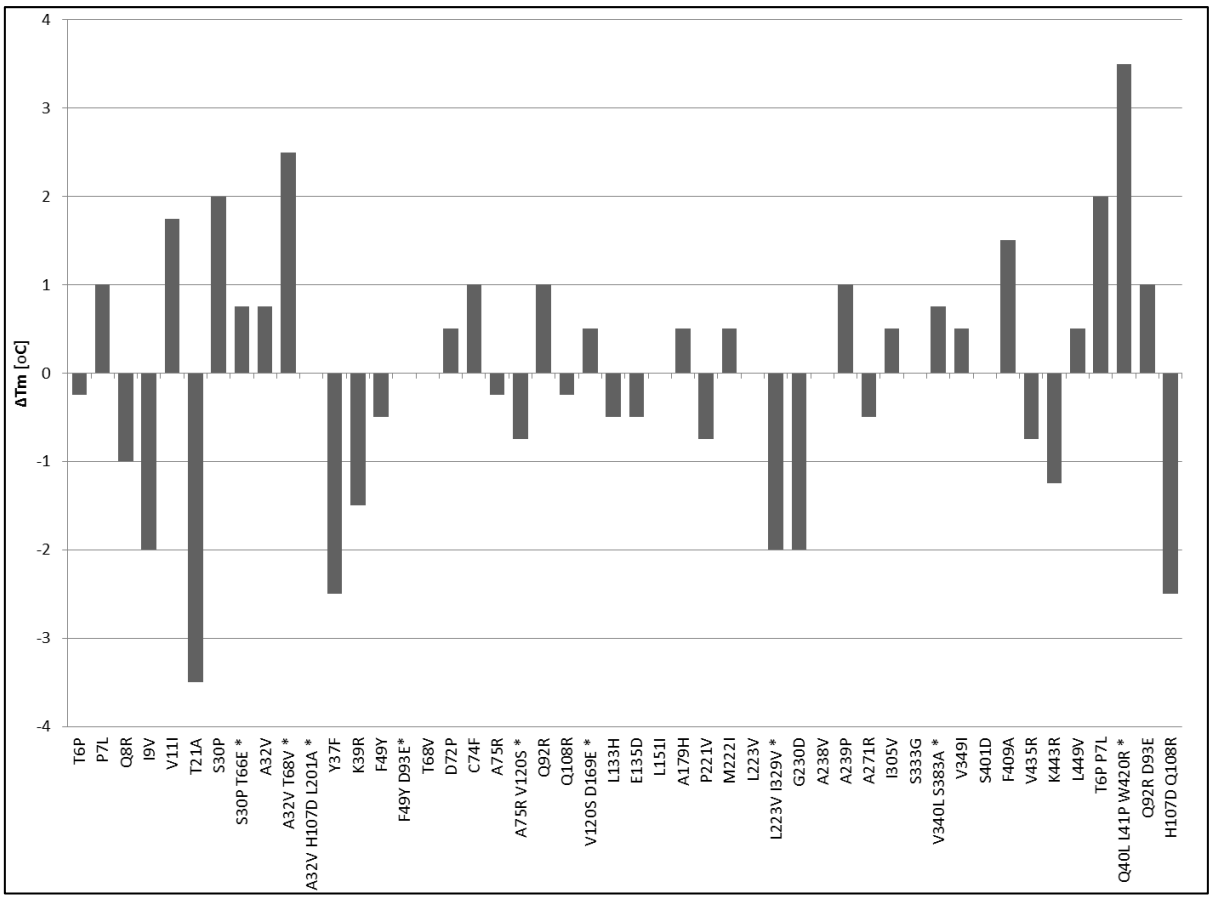


Figure 12: Melting temperatures of consensus bovine AdR variants relative to the wild type AdR (T_m=48°C). Variants market with '*' are not among the 88 proposed variants, and contain more than one back to consensus mutation.

Melting temperature (T_m)[k] of the wild type bovine AdR was 48°C, 21 tested variants showed 0.5°C or higher increase in T_m, 16 showed 0.5°C or higher decrease in T_m, and 11 variants showed no noticable difference in comparison to the WT (Figure 12 on the previous page). Out of the nine variants with more than one back to consensus mutation, five showed increased, two decreased and two unchanged T_m in comparison to the WT. In the initial results 44% of the back to consensus variants showed improved thermostability of AdR. The best single variant (S30P) improved thermostability 2°C and the best triple variant (Q40L+L41P+W420R) 3.5°C. The remaining variants and combinations of successful variants still need to be characterized to show the full potential of the described here approach.

To elucidate relationship between the sequence identities or origin of the sequences used for generation of consensus mutations and their effect on thermostability, the validated variants were grouped by sequence set and effect on stability (Table 9). Mutations originating from alignments with sequences 60-100% identical to the bovine AdR and from thermostable homologs in majority had positive effect on stability, whereas the mutations coming from sequence sets 30-60% and 30-100% identical to the bovine AdR showed mostly negative or no effect on the proteins thermostability. Thus, suggesting that thermophile and highly similar sequences are particularly suited for application in the sequence consensus apporach. Those results provide first insights into the unexplored relationship between the origin of the sequences used for consensus approach, and its effectiveness.

Table 9: Summary of tested variants effects on stability depending on the sequence set ('1': 30-60% sequence identity with the bovine AdR, '2': 30-100%, '3': 60-100%, '4': homologs from thermostable organisms).

Sequence set	Nr. of tested variants with mutation from the set	Positive effect on stability	No effect on thermostability	Negative effect on thermostability
1	32	28.1%	28.1%	43.8%
2	20	35.0%	20.0%	45.0%
3	13	69.2%	23.1%	7.7%
4	15	53.3%	33.3%	13.3%

Even though the thermostabilization efforts presented in this chapter were not performed on a CYP, the systematic consensus approach strategy presented here can be transferred to the

[k] Melting temperature (T_m) refers to a temperature at which free energy of proteins folded and unfolded states is equal, at such state half of the proteins are in folded and the other half are folded.

cytochrome P450 monooxygenases, or any other proteins for that matter. Therefore, the results presented here constitute one of the strategies for protein engineering of CYPs.

3.6 Modeling of CYP101A1 variants stereoselectivity towards methylated ethylbenzene derivatives

In this chapter, results of modeling of CYP101A1 variants stereoselectivity towards methylated ethylbenzene derivatives are presented. This work was done under supervision of Prof. Dr. Jürgen Pleiss, and in collaboration with Anja Eichler[l], Dr. Susanne Herter[l], Dr. Paul Kelly[l], Prof. Dr. Nicholas Turner[l] and Prof. Dr. Sabine Flitsch[l], who initiated the project, generated the CYP101A1 mutant library and performed screening against the substrate panel. The goal of this work was to elucidate molecular basis of stereoselectivity of the CYP101A1 library variants towards the methylated ethylbenzene derivatives. CYP101A1 from *Pseudomonas putida* is one of the most studied CYPs, and prototypic bacterial class I enzyme. Natively it accepts electrons from putidaredoxin and putidaredoxin reductase, but the variant used in this study was an artificial CYP101A1 fusion with CYP116B3 RhF reductase domain from *Rhodococcus sp.* strain NCIMB 9784.[110] The redox partner used in this artificial fusion system comes from a natural fusion system (class VII). Similarly to class I this system also contains a ferredoxin as a part of the electron transfer chain.

The library of CYP101A1 was semi-rationally designed, it combined described in the literature selectivity determining positions of this enzyme (Figure 13 on the following page).[111,112] The starting point for all variants in the library was variant Y96F of the CYP101A1-RhF fusion. The selectivity determining residues were divided into 7 sub-libraries (Figure 13), and subjected to site directed saturation mutagenesis. The sub-libraries targeted following sites: I:F87/F96, II:F98/T101, III:M184/T185, IV:L244/V247, V:G248/T252, VI:V295/D297, VII:I395/V396. All generated variants were screened using a colorimetric assay, in which indole was converted into blue indigo during 24h incubation. The screening identified 93 active variants.

[l] University of Manchester

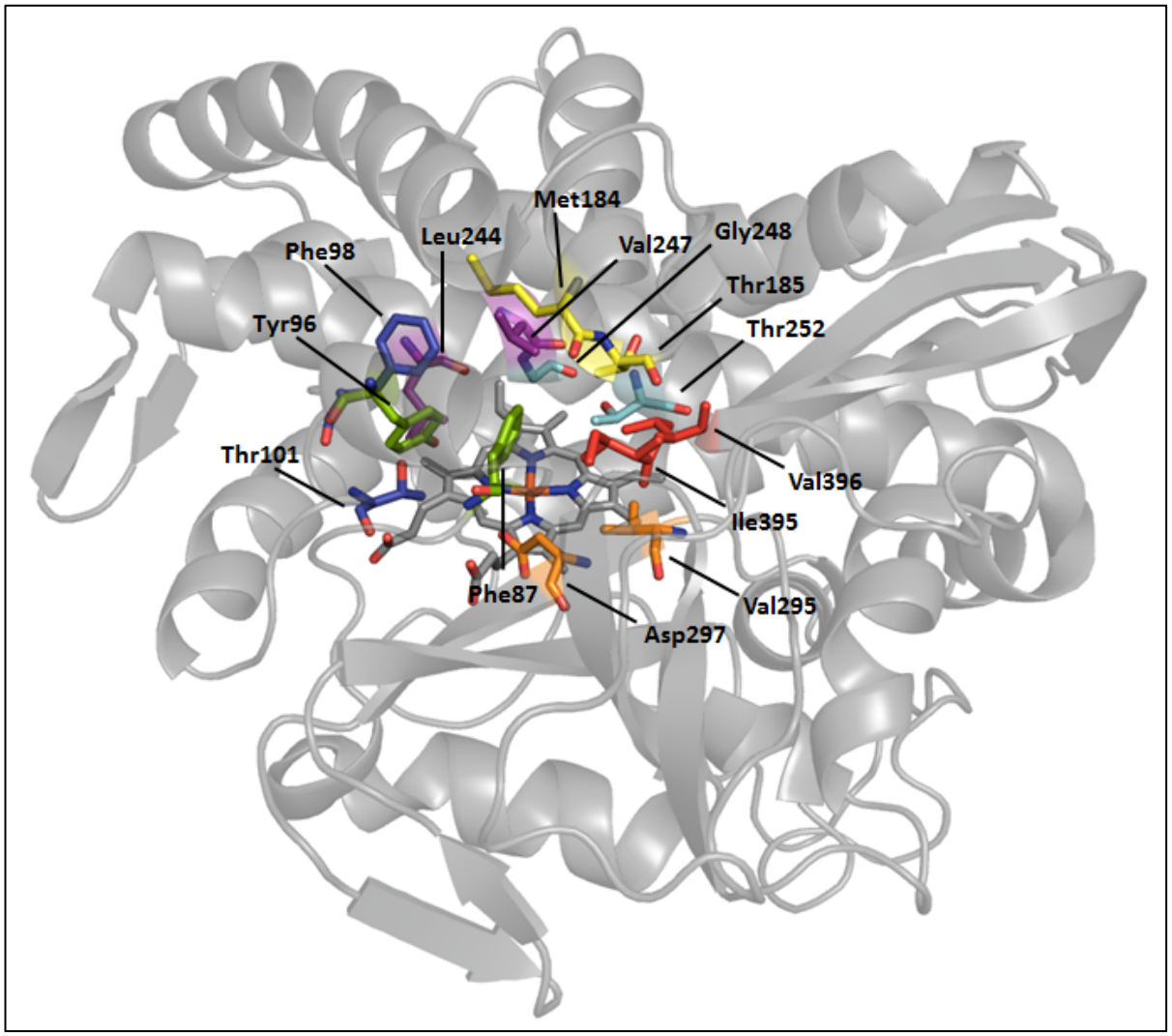


Figure 13. Selectivity determining positions of the CYP101A1 (PDB code: 1PHG).[38] The residues selected for generation of the highly enriched variant library are shown. The residues with the same color were subjected to saturation mutagenesis together, resulting in 7 sub-libraries targeting following residues: I: F87/F96, II: F98/T101, III: M184/T185, IV: L244/V247, V: G248/T252, VI: V295/D297, VII: I395/V396.

All identified active variants were subsequently subjected to screening by library pooling.[113] Each sub-library was divided into groups of 5-8 variant, and the groups were screened for conversion of 1-ethyl-4-methylbenzene. The results of pooling experiments as well as the literature[112] suggested the sub-libraries III and IV (targeting residues M184/T185 and L244/V247, respectively), as the most promising to obtain diverse selectivity. Variants from the two sub-libraries were screened against a panel of methylated ethylbenzene derivatives (Figure 14 on the following page): 1-ethyl-2-methylbenzene, 1-ethyl-3-methylbenzene and 1-ethyl-4-methylbenzene. The CYP101A1 variants performed oxidation of the substrates stereocenter hydrogens with the biggest product being alcohol, and traces of ketone. Enantiomeric excess values were calculated for all tested enzyme-substrate pairs (Table 10).

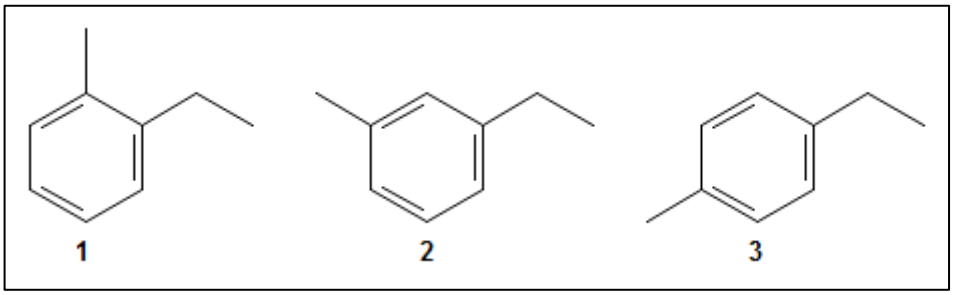


Figure 14. Studied substrate panel: 1: 1-ethyl-2-methylbenzene, 2: 1-ethyl-3-methylbenzene, 3: 1-ethyl-4-methylbenzene.

Table 10. Enantiomeric excess values of CYP101A1 variants in the conversion of methylated ethylbenzene derivatives. Variants selected for molecular dynamics simulations are in bold.

Variant	Enantiomeric excess [%]		
	1-ethyl-2-methylbenzene	1-ethyl-3-methylbenzene	1-ethyl-4-methylbenzene
Y96F	**60 (*S*)**	**7 (*S*)**	**35 (*S*)**
CV (Y96F/M184C/T185V)	51 (*S*)	12 (*S*)	7 (*S*)
VF (Y96F/M184V/T185F)	**64 (*S*)**	**74 (*S*)**	**28 (*S*)**
HI (Y96F /M184H/T185I)	53 (*S*)	14 (*S*)	29 (*S*)
FL_1 (Y96F/M184F/T185L)	58 (*S*)	30 (*S*)	27 (*S*)
LI (Y96F/M184L/T185I)	62 (*S*)	18 (*S*)	18 (*S*)
FV (Y96F/M184F/T185V)	62 (*S*)	15 (*S*)	26 (*S*)
CF (Y96F/M184C/T185F)	70 (*S*)	49 (*S*)	9 (*S*)
HF (Y96F/M184H/T185F)	62 (*S*)	54 (*S*)	15 (*S*)
FL (Y96F/L244F/V247L)	**>99 (*R*)**	**>99 (*R*)**	**43 (*R*)**
NL (Y96F/L244N/V247L)	**96 (*S*)**	**16 (*R*)**	**>99 (*S*)**
GL (Y96F/L244G/V247L)	77 (*S*)	60 (*R*)	71 (*S*)
CL (Y96F/L244C/V247L)	81 (*S*)	6 (*S*)	94 (*S*)
IL (Y96F/L244I/V247L)	n.d.	n.d.	35 (*S*)
SL (Y96F/L244S/V247L)	n.d.	n.d.	2 (*S*)

To elucidate molecular basis of stereoselectivity of selected CYP101A1 variants towards methylated ethylbenzene substrates ten 30 ns molecular dynamics (MD) simulations for each enzyme-substrate complex were performed (computational methods are described in chapter 6.1.1). Four variants with diverse selectivity were selected (Y96F, VF (Y96F/M184V/T185F), NL (Y96F/L244N/V247L) and FL (Y96F/L244F/V247L)) and simulated with the methylated ethylbenzene derivatives. Variant Y96F was selected because it was used as a template for all other variants. Variant VF was selected because it was the variant with the highest

stereoselectivity in library III. Variants NL and FL were selected because of the highest and complementary stereoselectivity amongst all tested variants.

Previously described method for estimation of CYP selectivity using molecular dynamics simulations was applied.[114–116] During the simulations, distances $d_{H\text{-}O}$ and angles $\alpha_{C\text{-}H\text{-}O}$ for the pro-*R* and pro-*S* hydrogens of the substrate were monitored to find near attack conformations (Figure 15). All frames of the simulation were assigned to pro-*R*, pro-*S* or non-active conformation based on distance and angle cut-offs. A conformation was considered to be near attack if for at least one of the stereocenter hydrogens, the distance $d_{H\text{-}O}$ was less than 0.35 nm and the angle $\alpha_{C\text{-}H\text{-}O}$ was 180±45°. The selectivity was calculated by summing up numbers of pro-*R* and pro-*S* near attack conformations from all simulations per enzyme-substrate complex and using them as equivalents of product formation.

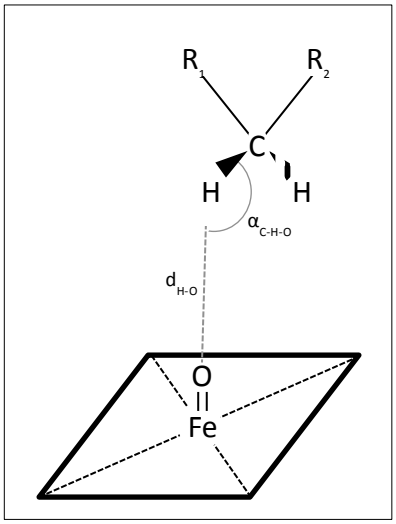


Figure 15: Schematic representation of basis for the prediction of selectivity: $d_{H\text{-}O}$ is the distance between the ferryl oxygen and hydrogen of the substrate, and $\alpha_{C\text{-}H\text{-}O}$ is the angle between ferryl oxygen, hydrogen and carbon of the substrate. A conformation was considered to be near-attack if $d_{H\text{-}O}$ was under 0.35 nm and $\alpha_{C\text{-}H\text{-}O}$ was 180±45°.

Comparison of the experimental enantiomeric excess values and values derived from the simulations (Table 11 on the following page), shows that the method allows for qualitative estimation of stereoselectivity. It is possible to distinguish between variants showing high *R*- and *S*-stereoselectivity, but precise prediction of stereoselectivity for all variant-substrate complexes was not possible. The estimated enantiomeric excess values deviated from the experimental results by 2% in the best case and 48% in the worst case, which is corresponding with 1-24% difference in product formation. In molecular dynamics simulations of CYP101A1 variants with ethylbenzene, the computationally estimated

enantiomeric excess values were opposite to the experimentally derived results. It was possible to identify the most selective variants but correct identification of *R*- and *S*-selectivity towards ethylbenzene was not possible.

Table 11: Comparison of enantiomeric excess values from experiments and values estimated based on the molecular dynamics simulations of selected CYP101A1 variants in conversion of methylated ethylbenzene derivatives.

	Enantiomeic excess [%]					
Variant	1-ethyl-2-methylbenzene		1-ethyl-3-methylbenzene		1-ethyl-4-methylbenzene	
	Experimental	Computational	Experimental	Computational	Experimental	Computational
Y96F	60(*S*)	35(*S*)	60(*S*)	15(*S*)	35(*S*)	60(*S*)
VF	64(*S*)	28(*S*)	16(*S*)	76(*S*)	28(*S*)	16(*S*)
FL	>99(*R*)	43(*R*)	38(*R*)	75(*R*)	43(*R*)	38(*R*)
NL	96(*S*)	>99(*S*)	75(*S*)	32(*S*)	>99(*S*)	75(*S*)

To evaluate if the simulations of the CYP101A1 variant-substrate complexes allow for extensive conformational sampling, heat maps showing the frequency with which the distinct distance and angle values are represented in the simulations were generated (Figure 16 on the following page and Figures S1-S23 on the pages 131-142). The heat maps showed that extensive sampling of the substrate conformations in the CYP101A1 binding pocket was achieved. This allowed identifying multiple binding modes of the substrates in the active site, and to explain the molecular basis of selectivity in the selected variants. Interestingly, the heat maps revealed that often many frequent conformations of the substrate were outside the near attack complex cut-offs, which suggests non-active binding modes. The frequent non-active binding modes might be the reason for low conversion of the studied substrates by the CYP101A1 variants.

To gain insight into the molecular basis of stereoselectivity of the studied CYP101A1 variants towards methylated ethylbenzene derivatives, all near attack conformations of the substrates from the molecular dynamics simulations were investigated.

The majority of experimentally tested variants showed highest stereoselectivity in the conversion of 1-ethyl-2-methylbenzene (Table 10 on page 51). Movement of this substrate in the active site of the variants, as well as conformational changes of the substrate itself, suggest that this overall high stereoselectivity is a result of the methyl group at the ortho-position limiting rotational movement of the ethyl group. Moreover, because of the compact size of the CYP101A1 binding pocket, the presence of methyl group at the ortho-position

makes the substrate wider, additionally restricts space for rotation of the ethyl group. Therefore, the methyl group at the ortho-position in 1-ethyl-2-methylbenzene provides steric restrictions resulting in an increased stereoselectivity in comparison to other substrates.

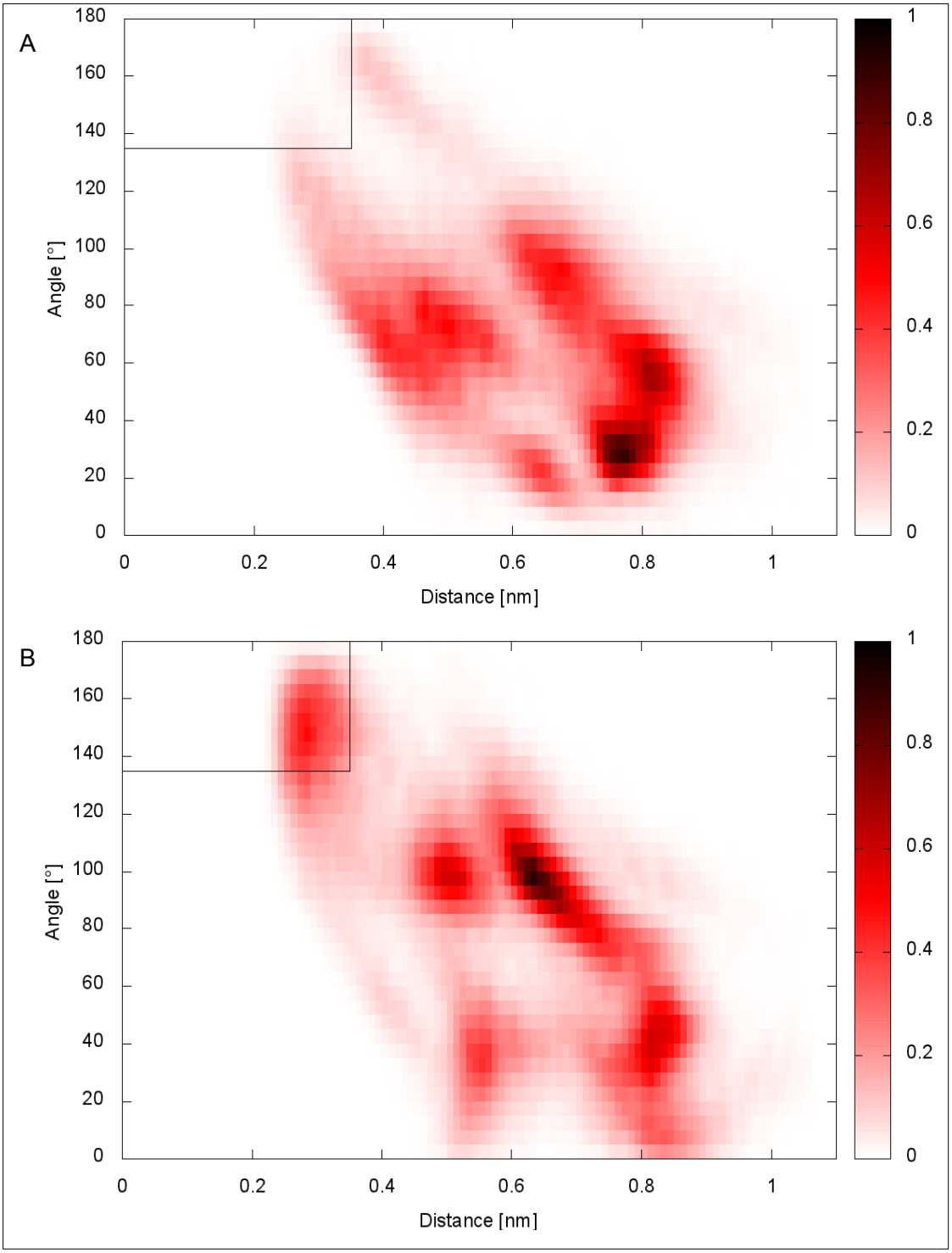


Figure 16. Heat map representing distance dH-O and angle αC-H-O values pro-*R* (A) and pro-*S* (B) orientations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant NL and 1-ethyl-2-methylbenzene (heat maps for all other variant-substrate pairs are available in the supplementary materials – chapter 6.1). The heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low (0) to high (1) number of simulation frames. The near attack conformations are inside the black frame (135-180° and 0-3,5 nm).

Variant Y96F is a parent variant for the whole library. The mutation in this variant increased size of the binding pocket and allowed for high mobility of the substrates, which resulted in multiple possible near attack conformations of the substrates. This can explain the relatively low stereoselectivity of the parent variant in comparison to the rest of the library. The near attack conformations of the substrates in the binding pocket were both with the benzene ring in a parallel and in a perpendicular orientation to the heme plane (Figure 17A on page 57). Nevertheless, the variant converted the substrates mostly to *S*-products, which was a result of a preferable stabilization of pro-*S* conformations resulting from a native shape of the CYP101A1 binding pocket.

Variant VF (Y96F/M184V/T185F) introduced an increased size of the residue at position 185 which is located over the heme and pushed the substrate closer to it. This change resulted in a more limited number of conformations in comparison to the variant Y96F. In majority of the near attack conformations the substrates were located in a way that the benzene ring was in a parallel orientation to the heme plane. Nevertheless, stereoselectivity towards 1-ethyl-2-methylbenzene and 1-ethy-4-methylbenzene did not change significantly, which is due to the fact that the pushing effect of the T185F mutation stabilized orientation of the benzene ring but did not influence in a significant way mobility of the ethyl group. An exception for this variant was 1-ethyl-3-methylbenzene which was converted in a significantly higher stereoselectivity in comparison to the Y96F variant: enantiomeric excess of 74% and 7%, respectively. The pro-*S* conformations of 1-ethyl-3-methylbenzene in the binding pocket of the VF variant were orientated exclusively in a way that the benzene ring was in a perpendicular orientation to the heme plane (Figure 17B on page 57) with the ethyl group stabilized in a small cavity close to the T252. Whereas, the pro-*R* orientations of the substrate were located with the benzene ring in a more parallel orientation to the heme plane (Figure 17B) and the ethyl group located in the same pocket. The higher frequency of pro-*S* conformations suggests that this is the preferred orientation of the substrate in the binding pocket.

While the library III variants introduced changes over the heme plane, the library IV variants introduced changes to the residues of the α-helix I resulting in a changed width of the binding pocket (Figure 17 on page 57). The FL variant showed the highest *R*-selectivity towards the methylated ethylbenzene derivatives. The mutations in this variant L244F/V247L significantly decreased the width of the binding pocket resulting in high stereoselectivity

towards 1-ethyl-2-methylbenzene and 1-ethyl-3-methylbenzene, and moderate stereoselectivity towards 1-ethyl-4-methylbenzene. In majority of the near attack conformations the benzene ring is in an almost perpendicular orientation to the heme plane. The high stereoselectivity is a result of the narrow binding pocket, which does not allow for high mobility of the substrate molecule and the orientation of the ethyl group which is close to the heme but pointing with its terminal carbon to the entrance of substrate access channel (residues F87 and F96). Such orientation of the substrates in the binding pocket is preferential for the formation of *R*-alcohols (Figure 17C on the following page). The moderate stereoselectivity towards 1-ethyl-4-methylbenzene can be explained by the fact that in the molecule with methyl-group at the para-position, conformations similar as with the other substrates are not possible. This is due to limited height of the binding pocket resulting from the V247L substitution (Fig 17C and D on the following page). The substrate appeared to be in a more parallel orientation to the heme plain, which increased space for rotational movement of the ethyl group, and therefore decreased stereoselectivity (Figure 17D).

Variant NL also decreases the size of the binding pocket. The variant NL showed the highest *S*-selectivity towards 1-ethyl-2-methylbenzene and 1-ethyl-4-methylbenzene, and low *S*-selectivity towards 1-ethyl-3-methylbenzene. This is a reverse in selectivity in comparison to the FL variant with only one amino acid substitution. In majority of the near attack conformations the benzene ring and ethyl group of the substrate are in a parallel orientation to the heme plane (Figure 17E on the following page). Those conformations of the substrate molecules are stabilized by L247 which is enforcing the pro-*S* orientation. The lower stereoselectivity towards 1-ethyl-3-methylbenzene is a result of lower stability of this molecule in the parallel orientation to the heme plane. The higher mobility of the substrate causes often perpendicular near attack conformations, which can result in both pro-*R* and pro-*S* products (Figure 17F on the following page).

The modeling of stereoselectivity in CYP101A1 towards methylated ethylbenzene derivatives, contributes a structural perspective to the strategies for protein engineering of CYPs. The method presented here could be used for qualitative *in silico* screening of CYP substrate libraries. Furthermore, the results highlight the fact that the CYP binding pockets are flexible and allow for multiple near attack conformations of substrates, but to achieve high selectivity it is necessary to as much as possible stabilize single conformation of the substrate in the binding pocket.

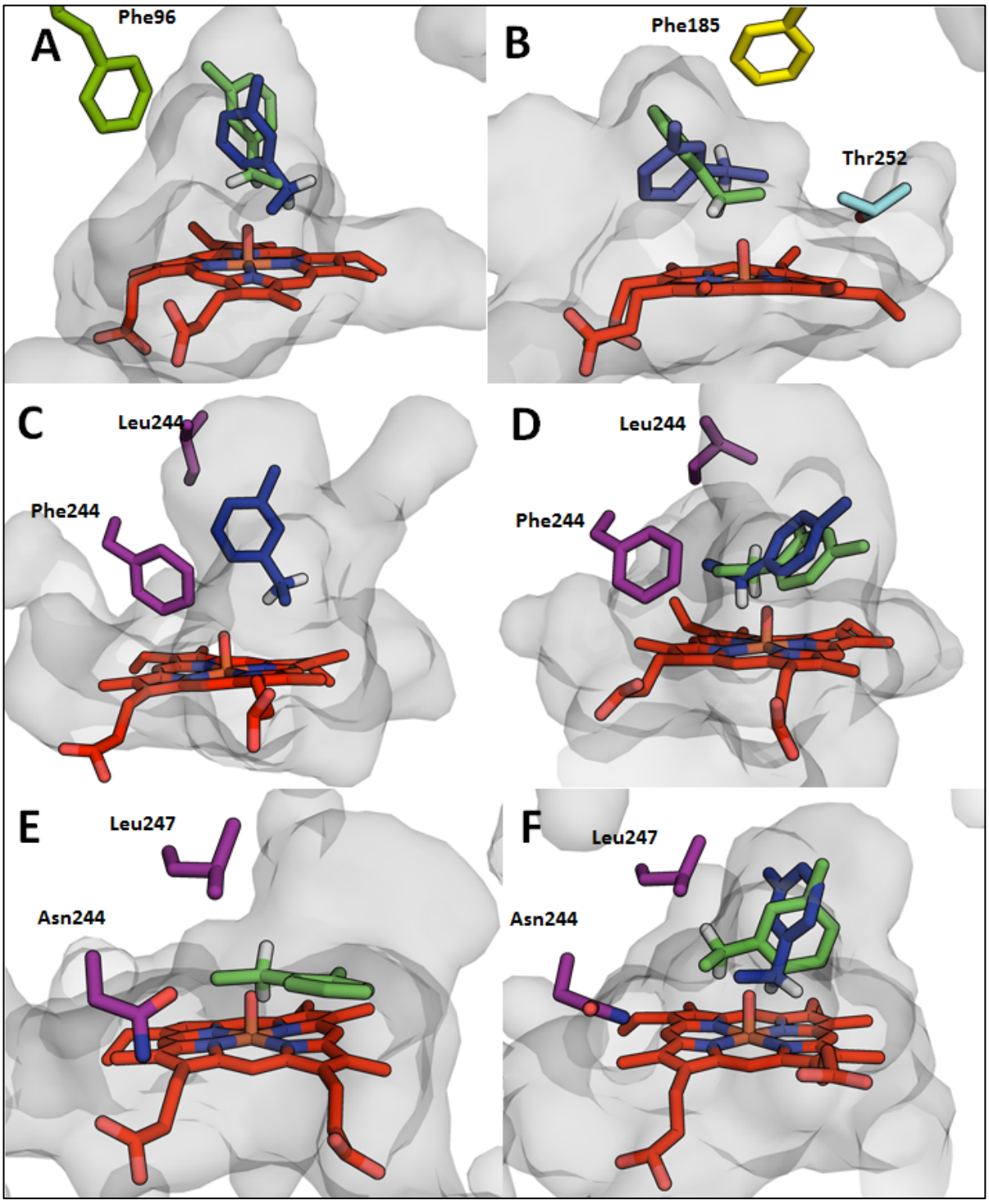


Figure 17: Examples of substrate orientations in the binding pocket of CYP101A1 variants, explaining trends in stereoselectivity towards methylated ethylbenzene derivatives. Heme compound I is colored red, substrates in the pro-*R* orientations are colored blue, whereas the pro-*S* orientations are colored green. A: Variant Y96F with 1-ethyl-3methylbenzene in pro-R and pro-S orientations. B: Variant VF with 1-ethyl-2-methylbenzene in pro-*R* and pro-*S* orientations. C: Variant FL with 1-ethyl-2-methyl in pro-*R* orientation. D: Variant FL with 1-ethyl-2-methylbenzene in pro-R and pro-S orientations. E: Variant NL with 1-ethyl-2-methylbenzene in pro-S orientation. F: Variant NL with 1-ethyl-2-methyl in pro-*R* and pro-*S* orientations.

4. Discussion

This chapter contains the discussion on the strategies for rational protein engineering of cytochrome P450 monooxygenase systems. The discussion covers following manuscripts: "Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance" (chapters 5.1),[31] "Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining" (chapters 5.2),[77] and "Redox partner interaction sites in cytochrome P450 monooxygenases: in silico analysis and experimental validation" (chapters 5.3),[81] specific discussion on those subjects can be found in the corresponding chapters. Additionally, it is complemented by discussion on three yet unpublished collaborative projects: the impact of linker length on cytochrome P450 monooxygenase fusion constructs (chapter 3.4), thermostabilization of bovine adrenodoxin reductase by sequence consensus approach (chapter 3.5) and modeling of stereoselectivity of CYP101A1 variants towards methylated ethylbenzene derivatives (chapter 3.6).

4.1 Strategies for rational protein engineering of cytochrome P450 monooxygenase systems

Protein engineering efforts concentrate around four major areas:

- stability,
- selectivity,
- specificity,
- activity.

Here rational approaches to protein engineering of CYPs targeting all those areas will be discussed. Engineering of proteins' **stability** is most often addressed by sequence consensus approach[106,117] and structure-based approaches.[57,118] Engineering of enzymes' **activity** can be an effect of changes in protein-ligand interactions and optimization of features associated with catalytic mechanisms. In case of CYPs this can mean improvement of electron transfer by

changes in CYP-redox partner interactions. Improvement of enzymes' **selectivity** and changes in its **specificity** are arguably the most common motivations behind protein engineering of CYPs. Those efforts can be targeted with sequence and structure-based approaches, both focusing on finding residues involved in interactions with or recognition of the substrate, and modifying them to residues improving those features.

Rational protein engineering methods can be divided into sequence and structure based. The sequence based approaches aim at finding functionally relevant positions by analyzing the amino acid sequence and comparing it to sets of homologous sequences. Whereas, the structure based approaches require proteins' crystal structure or a homology model, and are based on investigating the dynamics of a protein itself or its interactions with other proteins or ligands.

In this dissertation a set of technics and identified functionally relevant positions that can aid all of the above mentioned protein engineering areas are described. Data mining lies at the basis of this work and it allowed for distilling current knowledge from diverse research areas to provide strong pillars that can support protein engineering efforts.

4.1.1 Engineering of CYPs stability

Engineering of proteins' stability in high temperature or other non-natural conditions is an active area of research. This is due to the differences in conditions that enzymes are adapted to in living cells and extreme conditions of industrial processes. Furthermore, improved thermostability often results in longer life-time, shelf-life and higher temperature optimum.[119] Generation of thermostable variants is often addressed using the sequence consensus approach.[106] A key to successfully applying this method is the generation of a high quality sequence alignment. The class-specific standard numbering schemes for CYPs are based on the structure-guided sequence profiles, therefore are a perfect tool for generating high quality sequence alignments. Additionally, the Cytochrome P450 Engineering Database (www.CYPED.BioCatNet.de) is a comprehensive source of sequences, since it collects virtually all known CYPs. Hence, the numbering scheme-based alignment of sequences collected in the CYPED can provide a solid starting point for the consensus approach stabilization of CYPs.

Unfortunately, there are only handful published examples of natural or engineered thermostable CYPs. The major reasons for that might be: necessity for a thermostable redox partner and complex dynamics necessary for opening and closing of the F-G loop to allow substrate to enter and product to exit the binding pocket. The first discovered thermostable CYP119 from *Sulfolobus solfataricus* gave multiple insights into thermostabilization of CYPs, which can be used in design of stabilized variants.[120] There are also few described in the literature efforts aimed at increasing thermostability of CYPs, which all were based on directed evolution approaches rather than the rational methods.[121,122] Those examples of successful thermostabilization of CYPs, and the natural thermostable enzymes show that protein dynamics should not be prohibitive in the thermostabilization of CYPs. With the discovery of further natural thermostable CYPs also redox partners of those enzymes were found.[123,124] Even though, those proteins are not standard prokaryotic class I redox partners they share ferredoxin component which in principle could deliver electrons to other CYPs. Furthermore, there are efforts directed at generating thermostable redox partners. One example is described in this dissertation project, aiming at rational thermostabilization of the bovine adrenodoxin reductase (chapter 3.4) and adrenodoxin. Those redox partners could be used with thermostabilized mammalian class I or other compatible CYPs. Thermostabilization of adrenodoxin reductase shows utility of family specific databases in the design of thermostable enzymes, and paves the way for rational thermostabilization of CYPs. Hence, with a systematic approach to protein stabilization (as described in the chapters 3.5 and 4.2) it should be possible to improve stability of any CYP using rational protein engineering methods.

4.1.2 Engineering of CYPs selectivity and specificity

Selectivity and specificity are probably the most frequently targeted by protein engineering properties of CYPs. Continuous efforts in this area resulted in a wealth of variants with diverse substrate and product spectra. The rational strategies for protein engineering can be based on sequence and structure analyses. Based on the results presented in this dissertation two complementary approaches aiming at improving selectivity and specificity can be devised. The first one is based on incorporation of the universal selectivity determining positions into the CYP protein libraries, and the second one is based on insights from molecular dynamics simulations.

The substrate interacting regions of CYPs were identified and named substrate recognition sites (SRSs).[40] The SRSs cover the substrate access channels and binding pocket, therefore house the majority of positions influencing selectivity and specificity in all CYPs.[77] Since not all substrate recognition site positions directly interact with substrates, it was of interest to identify positions which are influencing selectivity in a wide range of CYPs, and can be used in combinatorial enzyme libraries of diverse CYPs. Previous sequence and structure analyses allowed identifying universal selectivity determining position in SRS5, which is corresponding to residue A328 in CYP102A1.[59] Subsequently this position along with another selectivity determining residue (F87) was selected as a basis for a minimal and highly enriched mutant library.[58,125] The variants exhibited highly diverse selectivity, specificity and improved activity. Such library combining only two selectivity determining positions of CYP102A1 showcased that it is possible to generate a highly diverse set of CYP variants with small number of rationally selected positions. This study shows potential benefit from identification of universal selectivity determining positions in CYPs.

A set of structurally corresponding positions influencing selectivity among a wide range of CYPs was identifies in the study presented in chapter 5.2. The sequence-based literature mining allowed for identifying the most frequently mentioned in the literature positions, which turned out to lie almost exclusively on the substrate recognition sites. Among 98 SRS, positions 14 were mentioned in the publications about more than 32 CYPs. Based on the literature information about those positions, it was proposed that they are most probably universal selectivity determining positions influencing selectivity of a wide range of CYPs towards diverse substrates. Since selectivity is mainly determined by the interactions between substrate and the enzyme, which in turn also influences specificity and activity, these positions will most probably also influence those properties. Both F87 and A328 of CYP102A1 are corresponding to the most and third most frequently described positions in all CYPs (Figure 9 on page 30), respectively. Remarkably, the universal selectivity determining positions are frequently appearing in protein libraries generated through both rational and directed evolution approaches. The best fatty acid ω-hydroxylating variants of CYP153A6 from *Marinobacter aquaeolei* are based on mutations of a pair of third most frequently mentioned positions: A307 (corresponding to class II standard position 264) and L357 (corresponding to class II standard position 328). Libraries resulted from directed evolution of CYP102A1 also often contain multiple universal selectivity determining positions.[126–128] The CYP101A1 library presented in the chapter 3.6 even though designed independently from the

study on identification of the universal selectivity determining positions contains almost exclusively those positions.

Incorporation of the structurally corresponding universal selectivity determining positions in any library of CYP variants should be beneficial, especially for proteins without known structure where it is not possible to pinpoint specific interactions between the enzyme and substrate. Transfer of the universal selectivity determining positions to any CYP sequence is possible thanks to the online numbering scheme tool available via www.CYPED.BioCatNet.de.

The studies of protein-substrate interactions using docking and molecular dynamics (MD) simulations provide many insights for engineering of proteins with known crystal structure or for which it is possible to generate a high quality homology model. In this dissertation the structure based approach utilizing MD simulations was used to estimate and understand the stereoselectivity of CYP101A1 variants towards methylated ethylbenzene derivatives (chapter 3.6). This study gave interesting insights into how the substrates behave in the CYP binding pocket. Generation of heat maps allowing for visualization of the mobility of the substrate during the MD simulation showed that even in the variants with high selectivity the substrates sample diverse conformations in the binding pocket. Only small fraction of those conformations (<15%), are near attack and could represent the reaction intermediate complex. Investigation of the near attack complexes revealed that for the enzymes exhibiting high selectivity, the number of possible conformations is significantly lower than for the enzymes with low selectivity, but still not limited to only one. The situation might be different in the natural substrate-enzyme complexes (e.g. CYP101A1 and camphor),[114] for which substrate fits the binding pocket in a single orientation as in the key and lock or hand and glove analogies.[129] The complementarity of substrate and enzyme binding pocket as in the natural complexes, should be the goal of protein engineering efforts aiming at generating highly specific enzymes.

4.1.3 Engineering of CYPs activity

Activity of an enzyme can be influenced by multiple factors, therefore different protein engineering approaches can result in improvement of CYPs activity. Optimization of the enzyme-substrate interaction besides influencing specificity and selectivity often also affects its activity. But in contrast to mutations influencing selectivity, which are mostly located in

the substrate access channels and the binding pocket, mutations influencing activity can be located in various parts of the protein.[130] Based on the results of this dissertation three approaches aimed at optimizing enzymes' activity can be devised. The first strategy involves maintaining conserved residues unaffected by mutations or bringing them back to the family consensus, the second is aimed at optimizing the linker region between CYP and redox partner in artificial fusion proteins and the third is aimed at improving the CYP-redox partner interactions.

The first approach is based on the assumption that the conserved residues are more crucial for maintaining proteins' activity and stability than the rest of its sequence. The conservation analysis results (Chapter 5.1), provide a list of conserved positions in class I and class II CYPs. This list can be used as a 'black list' of positions which should not be mutated, because their change will most probably negatively influence conserved functional or structural features of the enzyme.[36,76] Another use case of the list could be bringing the outlier proteins back to the family or class consensus. Such back to consensus mutation (G307A) increased activity of CYP153A6 from *M. aquerioli* towards different fatty acids 2- to 20-fold.[30]

The second approach is based on re-designing linker regions for fusion proteins of CYP and non-natural redox partners. This approach is important in case of CYPs for which the natural redox partners are not known, but also CYPs with known redox partners that are engineered to be fusion proteins. Generation of fusion proteins was shown to positively influence the CYPs activity and simplify the experimental setup and screening.[28,70] Therefore, in the study presented in chapter 3.4, a generic strategy for re-designing the linker region in artificial CYP-redox partner fusion systems was established. The strategy uses the natural linker of the redox partner as a template. By sequence analysis of the homologous proteins the amino acid conservation in the linker region is analyzed. In this approach the conserved positions should be kept unchanged whereas the rest of the linker is varied. The advantage of this approach is keeping the possible conserved interactions between the linker region and the redox partner unchanged, and maintaining the natural flexibility of the linker region. The flexibility of linker regions was often disturbed by inclusion of GGS peptide sequences into the artificial fusion constructs, whereas the natural linker sequences usually do not contain many glycine residues.[28] The approach presented in this dissertation was developed and tested on a fusion of CYP153A6 from *Marinobacter aquaeolei* and the redox partner of natural fusion protein CYP116B3 from *Rhodococcus ruber*. The best variant had a linker region only two amino acids shorter than the natural linker and showed 50% improved expression and 1.7 fold

improved conversion of dodecanoic acid relative to the CYP concentration. This approach can be used with any CYP116B family redox partners, exactly as described in chapter 3.4, or additional sequence conservation analysis should be performed for fusion proteins with redox partners from other CYP homologous families.

The third approach is based on improvement of CYP-redox partner interactions. This strategy is also beneficial for non-native CYP-redox partner systems. This approach utilizes the identified here CYP redox partner interaction sites (RPISs) (chapters 5.3). Crystal structures of CYP-redox partner complexes suggest that the interaction sites among different CYPs and redox partners are most probably overlapping.[43,45,46,74] Through sequence, structure and literature analyses it was possible to identify six RPIS. Positions described in the literature to influence the CYP-redox partner interactions are often structurally corresponding between diverse CYPs suggesting interaction hot spots on the proximal surface, which probably influence the interaction in most of the CYPs. To test this assumption a generic strategy requiring minimal screening efforts was designed to improve the CYP-redox partner interaction based on the RPIS mutations. The strategy in the presented implementation was targeting a fusion protein of CYP153A6 with redox partner domains of CYP102A1.[30] Residues at the RPIS positions influencing the CYP-redox partner interactions in mammalian CYPs were compared between CYP153A6 and natural heme domain of CYP102A1. Most of the positions differing in charge between the two proteins were selected for mutagenesis, which resulted in six variants. Five variants were successfully expressed and screened for activity and electron coupling. The best designed variant showed increase of electron coupling from 68% to 89%, but decrease in catalytic activity. The approach can be extended in several ways to increase probability of finding variants with more dramatically improved electron coupling and catalytic activity. One way of extending the strategy could be to create RPISs variants mimicking all residues described to influence the interaction, not only the charged ones. Some of the described in the literature variants with improved activity based on CYP-redox partner interactions did not influence surface charge.[131,132] Another way of modifying the approach would be to mutate positions influencing the interaction to the most frequently found residues in the whole protein family. Finally a semi-rational approach, involving site-saturation mutagenesis of the RPIS positions, which would require the biggest screening efforts but could have the highest chances for finding the most active variant. Additionally, use of semi-rational approaches would significantly expand current knowledge of the CYP-redox partner interactions by increasing number of tested RPIS positions.

4.2 Insights into the sequence consensus approach for protein thermostabilization

Adrenodoxin reductase (AdR) is part of the mitochondrial class I redox systems in mitochondrial cytochrome P450 monooxygenases, where the electrons are transferred from NADPH through AdR and adrenodoxin (adx) to the CYP.[26] The thermostability of bovine adx was recently increased by 15 °C (Kembaren, R. F., Wijma, H. J. & Janssen, D. B. Unpublished data), here the initial results of thermostabilization of bovine AdR are discussed.

Thermostabilization of adrenodoxin reductase is challenging because of its cofactors (NADPH and FAD), interactions with adrenodoxin, and conformational changes necessary for the electron transfer.[104,105] To keep activity the variants must not disturb any of those features. Since sequence consensus approach for protein stabilization is based on comparison of the protein to its homologs it is not expected to introduce changes inactivating the protein. Additional advantage of this approach in comparison to other computational or directed evolution based approaches is a limited number of variants necessary for screening.[57] This is especially important in large proteins like the 461 amino acid long AdR.

Consensus approach strategy used in this study was extended to define relationship between the origin of sequences used for generation of sequence alignments and effectiveness of the predicted mutations. Hence, four groups of AdR homologs were generated: proteins 30-60%, 60-100% and 30-100% identical to the bovine AdR and thermostable homologs. There were 97 back to consensus mutations suggested. To minimize the screening efforts, the structures of all mutants were subjected to molecular dynamics simulations. 84 out of the 97 point mutations suggested by the sequence consensus, did not increase protein flexibility in molecular dynamics simulations[57] and therefore were selected for experimental characterization.

To rationalize the effects of mutations on the AdR stability, the specific residues were localized on the AdR crystal structure (PDB code: 1E1M[105]). Three of the 84 positions selected for mutagenesis interact with the cofactors. Residue K39 is located close to adenine rings of the FAD whereas L151 and I329 are close to the adenine rings of NADPH. Variant K39R and L223V/I329V (variant L223V alone did not influence melting temperature (T_m)) decreased proteins T_m by 1.5°C and 2°C, respectively. Whereas, variant L151I did not show any influence on the T_m. This shows that binding of the cofactors is very sensitive and even

seemingly small changes of it can have negative effects on the proteins stability and possibly activity. Variant Q40L/L41P/W420R showed the highest improvement of T_m (3.5°C). Mutations Q40L and L41P located on a loop were selected to be combined because of their proximity on the sequence and structure. However, W420R is possibly responsible for the high improvement of T_m in this variant, because of introduction of surface accessible charge instead of tryptophan. R420 could also establish a salt bridge with D415. Salt bridges are described to significantly improve protein stability.[133,134] Variant T21A showed the highest decrease of T_m (3.5°C). This substitution removed a hydrogen bond between T21 and S365. Hydrogen bonds are also known to have a significant impact on the protein stability.[135] With screening of the remaining variants and combination of stabilizing mutations, it is expected to find a variant with even more significantly improved T_m. Additionally activity measurements of the best variants have to be performed to test their influence on the proteins functioning.

Besides finding the most stable variant of the bovine AdR this study was aimed at exploring the relationship between effectiveness of alignment in the sequence consensus approach and sequence identity and origin of the proteins used for its generation. Even though the consensus approach has been extensively used for protein stabilization,[106] there are no studies systematically analyzing this relationship, and providing rules for generation of the most effective alignments for application in protein stabilization. Therefore sequence alignments based on four sequence sets were used to find consensus variants: sequences 30-60%, 30-100% and 60-100% identical to the bovine adrenodoxin reductase and sequences of its thermostable homologs. Based on the current results (Table 9 on page 48) the alignments of sequences the most similar to the bovine adrenodoxin (60-100%) and the thermostable homologues were the most effective (highest percentage of variants improving the T_m). But interestingly, the best variant Q40L/L41P/W420R includes consensus mutations originating in all four sequence sets (Q40L/L41P from 30-60%, 30-100% and thermostable and W420R from 60-100%), and the second best variant A32V/T68V originated from the sequence sets with the lowest percentage of variants improving the T_m (A32V from 30-60% and 30-100%, and T68V from 30-60%). Those results show that even though mutations from certain sequence sets are more probable to have positive effect, variants originating from all sequence sets can significantly improve stability of a protein.

The experimental characterization of the designed variants is not complete and the final conclusions will be formulated after all variants have been tested. Nevertheless, the initial results already show interesting trends suggesting that the use of multiple alignments for the

consensus approach can be beneficial, while still allows for keeping the number of variants necessary for testing relatively low.

4.3 Molecular basis of CYP101A1 stereoselectivity towards methylated ethylbenzene derivatives

Combinatorial protein libraries generated using both rational design[54,58,136] and directed evolution approaches[51,137,138] allow to unlock new substrate and product spectra of the natural enzymes. Protein library presented in the chapter 3.6 shows a great potential of combining rationally selected positions with site-saturation mutagenesis. Such approach allowed for identifying a diverse set of CYP101A1 variants with improved stereoselectivity towards methylated ethylbenzene derivatives. 12 out of 14 positions selected for generation of the library are also included in the list of the universal selectivity determining positions, residues: F87, T101, M184, T185, L244, V247, G248, T252, V295, N297, I396, V396 are structurally corresponding to class II standard positions: 73, 87, 180, 181, 260, 263, 264, 268, 328, 331, 437, 438. Those residues were chosen for mutagenesis because of the described high importance in influencing selectivity of CYP101A1.[111,112] The identification of universal selectivity determining positions by systematic literature mining and generation of CYP101A1 library were independent projects, which shows great potential of including those positions in combinatorial protein libraries of other CYPs. The library allowed identifying variants with high *R*- and *S*-stereoselectivity for almost all screened substrates: 1-ethyl-2-methylbenzene >99% R and 96% S with variants FL and NL, 1-ethyl-2-methylbenzene >99% R and 76% S with variants FL and VF and 1-ethyl-4-methylbenzene >99% S with variant FL.

To understand the molecular basis of stereoselectivity of the selected library variants molecular dynamics simulations of variant-substrate complexes were performed. Previously described method for estimation of CYP selectivity was used to analyze the MD simulations.[114–116] The method is based on monitoring the distance between the heme compound I ferryl oxygen and substrate hydrogen atom and the ferryl oxygen-substrate hydrogen-substrate carbon angle in all simulation frames (Figure 15 on page 52). The near attack conformations are identified in the MD simulation if for at least one of the stereocenter hydrogen atoms the distance is <3.5 nm and angle 180±45°. Those criteria applied for studied here set of variants and substrates allowed for qualitative estimation of stereoselectivity

without introducing bias to the simulations. Additionally the near attack conformations of the substrates were used to elucidate molecular basis of stereoselectivity in the selected variants. The differences between estimated and experimental enantiomeric excess values are expected to be a results of an inversion of the configuration.[139] This phenomenon can be described as a change in substrate orientation between the hydrogen abstraction and product formation, which may result in abstraction of pro-*R* hydrogen and formation of *S*-product. Since, the approach is based on the identification of near attack conformations and allows for modeling of hydrogen abstraction, it does not allow taking inversion of configuration into account.

The method used in this study as well as its variations taking into account only distance criterion was successfully applied to model selectivity of multiple CYPs towards diverse substrates. The initial studies by Paulsen et al.[114,116] on modelling of selectivity were based on sub-nanosecond simulations of wild type CYP101A1 with camphor and camphor derivatives, starting from an X-ray structure of a CYP101A1-camphor complex.[114,116] Here, length and number of the simulations were extended in comparison to the previous studies, to allow for conformational changes induced by the amino acid substitutions or presence of the substrate, and extensive sampling of the conformations of the substrate as well as the protein. Sufficient sampling of the conformational space of the CYP-substrate complex is a crucial point in application of the MD methods for modeling of selectivity. This issue becomes more significant when there is no crystal structure of the analyzed enzyme-substrate complex, and the substrate has to be artificially placed in the binding pocket or substitutions of amino acids have to be introduced.[140,141] In such situation, conformations of both substrate and protein need to adjust, and sub-nanosecond simulations might not be sufficient.

Besides using this method for explanation of selectivity it could be also used for virtual screening of protein libraries. Application of virtual screening strategies based on molecular docking[142] and molecular dynamics simulation[114,143–146] show high potential in decreasing the amount of high-throughput screening necessary to identify variants with improved selectivity. Additionally virtual screening methods could be applied to validate quality of homology models. Current state of the art tools like QMEAN[147] or ANOLEA[148] allow to assess geometrical correctness of a model, but do not allow to successfully distinguish between two structures with perfect geometry but in non-productive and productive conformations. Molecular dynamics simulation-based structure assessment however, could allow identifying proteins in productive and non-productive conformations, and provide a better starting point for structure-based protein engineering.

4.4 Conclusions

Cytochrome P450 monooxygenases have widely recognized high potential for commercial applications, but due to limitations caused by low stability, not always optimal specificity, selectivity and activity, the industrial use of those enzymes remains challenging. Numerous successful attempts at improving the properties of CYPs show high potential of protein engineering in overcoming those shortcomings to allow for economically valid processes. In this dissertation novel tools and systematic sequence, structure and literature analyses, supporting rational protein engineering are presented. Thanks to the standard numbering schemes it is now possible to communicate about mutations and amino acid positions with use of unified standard position numbers, and to reliably transfer functional information between different members of the family. The identified here universal selectivity determining positions in combination with the standard numbering schemes, can help in limiting the number of potential mutagenesis targets, and therefore accelerate identification of variants with promising substrate and product spectra. Since, majority of CYPs require external redox partners and are often engineered to be artificial fusion proteins, the identification of redox partner interactions sites and the strategy to optimize linker regions in such fusion constructs, can help in overcoming the issues with CYP activity and electron uncoupling. Finally, a systematic approach to thermostabilization of CYPs and their redox partners based on the family specific protein databases should allow for overcoming the stability issues of CYPs.

All of the herein mentioned novel approaches and insights into CYP protein engineering were possible to attain thanks to looking at a big picture of the protein family. Data mining approaches and generalized view on the protein family and its properties, might not yet give a single amino acid substitution answer to an issue, but can provide useful strategies to be applied in solving issues in enormous numbers of homologous proteins.

5. Publications

5.1 Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance

5.1.1 Abstract

Cytochrome P450 monooxygenases (CYPs) constitute an ubiquitous, highly divergent protein family. Nevertheless, all CYPs share a common fold and conserved catalytic machinery. Based on the redox partner type, 10 classes of CYPs have been described, but most CYPs are members of class I accepting electrons from ferredoxin which is being reduced by FAD-containing reductase, or class II accepting electrons from FAD- and FMN-containing CPR-type reductase. Because of the low sequence conservation inside the two classes, the conserved class-specific positions are expected to be involved in aspects of electron transfer that are specific to the two types of redox partners. In this work we present results from a conservation analysis of 16732 CYP sequences derived from an updated version of the Cytochrome P450 Engineering Database (CYPED), using two class-specific numbering schemes. While no position was conserved on the distal, substrate-binding surface of the CYPs, several class-specific residues were found on the proximal, redox partner-interacting surface; two class I-specific residues that were negatively charged, and three class II-specific residues that were aromatic or charged. The class-specific conservation of glycine and proline residues in the cysteine pocket indicates that there are class-specific differences in the flexibility of this element. Four heme-interacting arginines were conserved differently in each class, and a class-specific substitution of a heme-interacting tyrosine by histidine was found, pointing to a link between heme stabilization and the redox partner type.

5.1.2 Introduction

Cytochrome P450 monooxygenases (CYPs) constitute a large, diverse, and ubiquitous family of heme-containing enzymes. The number of known CYP genes is approaching 20000.[149] Sequences that share more than 40 % of identity are clustered into families, sequences with more than 55 % of identity into subfamilies according to the standard CYP nomenclature.[150] In the Cytochrome P450 Engineering Database[151] (CYPED), those families and subfamilies are referred to as superfamilies and homologous families, respectively. CYPs catalyze the oxidation of a broad spectrum of substrates.[152] Because of their crucial role in drug metabolism and their possible synthetic applications,[10,153,154] there is a growing interest in understanding the relationship between sequence, structure, and function of CYPs. Most of the CYPs require electrons from an external redox partner for their catalytic activity. Based on the redox partner type, CYPs were classified into 10 different classes,[26] where class I and class II enzymes account for more than 90 % of all CYPs. Class I consists of CYPs that accept electrons from ferredoxin-type proteins (with an iron-sulfur cluster as cofactor) which shuttle electrons from NAD(P)H-dependent ferredoxin reductase to the heme domain. Class II is the most common and divergent group of CYPs. All members of this class accept electrons from FAD- and FMN-containing CPR-type reductases. One of the bottlenecks which hinder CYPs from being widely used in synthesis is their redox partner-dependence,[10] because suboptimal CYP-redox partner interactions results in low electron transfer rates and therefore low catalytic activity. It has been suggested that electrostatic interactions are the basis of the redox partner recognition, with positively charged residues on the proximal surface of the heme domain interacting with negatively charged residues on the surface of the electron donor.[155–157] However, it is not known whether there are significant differences between the two classes on the sequence level that could explain the redox partner preferences.

Since CYPs are a highly diverse family, the sequence identities between distant enzymes might drop below 15 %. However, all CYPs have a highly similar structure. A general naming system for structural elements was proposed based on the first crystal structure of the CYP from *Pseudomonas putida*: α-helices A-L, β-strands 1-5, the meander loop, and the cysteine pocket.[38] In addition, six substrate recognition sites were identified.[38] The proximal side of the heme domain, where the heme is close to the protein surface, is more conserved in structure, whereas the distal part of the heme domain, which is involved in substrate recognition, is more variable.[158] Due to low sequence identity within the CYP family, a systematic sequence comparison is not straightforward, because it depends on a reliable multisequence alignment.

Recently, a generic strategy of establishing a standard numbering scheme for large families of proteins was proposed.[71] A standard numbering scheme of a large protein family consists of a family-specific sequence profile generated from a set of representative protein family members. A numbering scheme for all members of a protein family is derived by profile-guided alignments of all protein sequences against the family-specific sequence profile and by transferring the internal residue numbers of a dedicated reference sequence to the respective positions in all sequences. As a result, each position of each sequence of a protein family is unambiguously numbered according to the reference sequence. As shown for metallo-β-lactamases,[159,160] antibodies,[161,162] and the ThDP-dependent decarboxylases,[71] unambiguous sequence numbering helps to analyze conserved, functionally or structurally relevant positions and to communicate about interesting protein variants.

The goal of this work was to establish a standard numbering scheme for CYP sequences contained in the CYPED,[151] and to perform a systematic analysis of sequence, structure, and function. Because there are significant structural and functional differences between the two classes of CYPs,[72,163] two separate numbering schemes were established. The numbering schemes help to compare the results from mutation studies of different proteins, but also provide a reliable tool for sequence alignments of proteins with low sequence similarity. Such sequence alignments were used to analyze class I and class II CYPs and to find the class-specific residues as well as residues conserved in all CYPs.

5.1.3 Methods

Cytochrome P450 Engineering Database

The Cytochrome P450 Engineering Database (CYPED)[151] was updated using one seed sequence of each homologous family. The seed sequences were derived from the last online version of the database. For each of the seed sequences, a BLAST[164] search was performed against the non-redundant protein database of the NCBI GenBank[165] using an E-value of 10^{-100} as cut-off criterion. Each hit was compared against all so far collected sequences in the database in order to classify the proteins according to their sequence similarity. Sequences with a global sequence similarity above 98 % were assigned to one protein entry, sequences with a sequence similarity above 55 % were grouped into homologous families and with above 40 % similarity into superfamilies. All newly found sequences were allocated to the predefined homologous families represented by the 626 seed sequences. From this dataset,

family-specific multisequence alignments and HMM profiles were generated using CLUSTALW[166] and the HMMER 3.0[167] toolbox, respectively. Partial sequences (shorter than 300 amino acids or annotated as fragment or partial sequence in the GenBank[165]) and sequences longer than 1500 amino acids were excluded from the database update. Sequences and structures for implementation and the application of the standard numbering scheme were derived from the updated version of the CYPED.

Class assignment

CYP sequences collected in the database were assigned to classes based on the literature data about the electron donor type.[26] Sequences with missing information about the electron donor type were assigned to classes based on their source organisms and their sequence similarity to characterized CYPs. CYPs from class III, IV, V, VI, IX and X were manually assigned based on the literature data.[26] Fusion proteins were assigned to classes VII (also known as IV) and VIII (also known as III) based on the fused redox partner type. The remaining CYPs were assigned to class I or class II based on the CYP nomenclature. Class I are in general bacterial CYPs which were assigned to superfamilies 51-70 and 101-299, and class II eukaryotic CYPs which were assigned to superfamilies 1-49, 71-99 and 301-750. As an exception, superfamilies containing mitochondrial CYPs were classified as class I based on their electron donor type. Although this class assignment strategy allowed for a minimization of wrong assignments, most of the sequences were not experimentally validated. Therefore, we expect that some non-typical CYPs might need reassignment in the future.

Reference alignments and position number assignment

The higher structural and functional similarity inside class I and class II was previously reported to considerably influence the quality of structural and functional modeling of CYPs.[163] Therefore, two separate numbering schemes were developed for class I and class II CYPs to allow for an accurate standard numbering. In order to provide a reliable and consistent method for the assignment of standard numbers to each CYP sequence, a method using family-specific reference sequences was applied.[71] For each of the classes I and II, a set of representative CYPs with available crystal structure (Table 12 on the following page) was chosen.

Table 12: Set of representative CYPs (24 class I, 14 class II) used for generating the class-specific sequence profiles based on structure alignments. The reference sequences (CYP101A1 for class I and 102A1 for class II) are in bold.

Class I			Class II		
PDB code	Organism	CYP	PDB code	Organism	CYP
3N9Y	*H. sapiens*	11A1	2HI4	*H. sapiens*	1A2
3K9V	*R. norvegicus*	24A1	3PM0	*H. sapiens*	1B1
1PHG	***P. putida***	**101A1**	1Z10	*H. sapiens*	2A6
3OFT	*N. aromaticivorans*	101C1	3KW4	*H. sapiens*	2B4
3LXI	*N. aromaticivorans*	101D1	1NR6	*O. cuniculus*	2C5
2Z36	*N. recticatena*	105	3TBG	*H. sapiens*	2D6
3ABA	*S. avermitilis*	105P1	3E6I	*H. sapiens*	2E1
1JIO	*S. erythraea*	107A	3CZH	*H. sapiens*	2R1
2Y46	*M. griseorubida*	107E	3NXU	*H. sapiens*	3A4
2WI9	*S. venezuelae*	107L	3SN5	*H. sapiens*	7A1
2JJO	*S. erythraea*	113A	3RUK	*H. sapiens*	17A1
2VE3	*Synechocystis sp.*	120A1	3EQM	*H. sapiens*	19A1
2IJ7	*M. tuberculosis*	121	3MDM	*H. sapiens*	46A1
2WM4	*M. tuberculosis*	124	**1ZOA**	***B. megaterium***	**102A1**
2XKR	*M. tuberculosis*	142			
1ODO	*S. coelicolor*	154A1			
1S1F	*S. coelicolor*	158A2			
2XBK	*S. natalensis*	161A			
3R9C	*M. smegmatis*	164A2			
1LFK	*A. orientalis*	165B3			
1UED	*A. orientalis*	165C			
1Q5D	*S. cellulosum*	167A1			
1N97	*R. thermophillus*	175A1			
4DNJ	*R. palustris*	199A2			

The representative CYPs were structurally aligned using the structural alignment tool STAMP.[94] These class-specific reference alignments were manually optimized to decrease the number of gaps. The curated alignments were subsequently used to generate the reference profiles using the *hmmbuild* program from the HMMER 3.0 software package.[167] From each class, the sequence of a well-investigated CYP was chosen as reference sequence for the numbering: CYP101A1 ($P450_{cam}$) from *Pseudomonas putida* for the class I numbering scheme, CYP102A1 ($P450_{BM3}$) from *Bacillus megaterium* for class II. CYP102A1 is not a class II CYP in a strict sense due to the bacterial origin, but the enzyme accepts electrons from a diflavin reductase and is structurally more similar to human than to bacterial CYPs, including the length of the meander loop and αJ/J'.[19] The two reference proteins are among the most studied CYPs. The standard numbering of each CYP was calculated by aligning its sequence to the respective reference profile. From each of the resulting alignments, a pairwise alignment of the respective query sequence and the reference sequence was extracted. Subsequently, the residue numbers of the reference sequence were transferred to the

respective positions of the query sequence (Figure S24 on page 146). Since the alignment of the query sequence to the specific reference sequence is guided by a structure-based profile, the assigned standard numbers indicate equivalent backbone positions in their structures (Figure S25 on page 147). This allows for the annotation of functionally relevant positions and regions as well as secondary structure elements (such as the α-helices A-L, the β-strands 1-4, the meander loop, and the cysteine pocket), based on knowledge of the respective positions in the two reference sequences. CYPs from classes I, III, IV, V, VI, and VII (class VII is also known as class IV) were annotated using the class I numbering scheme, because they accept iron-sulfur cluster proteins as an electron transfer protein. The class VIII CYPs (also known as class III) are chimeric proteins consisting of the heme and the CPR domain and class II CYPs were annotated using the class II numbering scheme. NADH-dependent CYPs (class IX) and independent CYPs which do not need a redox partner and use an intramolecular electron transfer system (class X) were annotated using the class II numbering scheme, based on their higher structural similarity in loop regions to class II CYPs. The standard numbers for all positions were included into the sequence view of the database web interface.

Web interface

A web application was integrated into the CYPED to allow users to apply the standard numbering scheme to their query sequence. Upon submission of a query sequence (Figure S26A on page 148), a search using the UBLAST method of USEARCH[164] against all sequences in the database is performed. Sequences without matches in the database (E-value above 10^{-10}) are rejected from the number assignment process, since they are too different from currently known CYPs. The query sequence is assigned to class I or II depending on the class assignment of the best matching sequence. The query sequence is then aligned against the class profile which is specific for the best matching sequence, and the standard numbers are assigned based on the alignment to the respective reference sequence (Figure S26B on page 148).

Selection of representative proteins

In order to cover the whole sequence space of the CYPs in the reference alignments of class I and class II, representative enzymes from different homologous families were chosen. Therefore, the sequences derived from the representative CYP structures of the two classes were separately clustered using USEARCH[164] into subfamilies by the criterion of intra-family

sequence identity above 50 %. The structures of class I and class II CYPs were sorted into 31 and 14 clusters, respectively. In order to ensure that all representative proteins were crystalized in their active conformation, only structures with substrate bound in the active site were considered. From each cluster containing at least one structure with bound substrate, the structure with the highest resolution was chosen as a representative protein for the development of the numbering scheme. This resulted in 26 structures representing class I and 14 structures representing class II (Table 12 on page 74).

Multisequence alignments using the numbering scheme

By aligning each sequence of the class I and the class II CYPs to the respective class-specific reference profile, multisequence alignments of class I and class II CYPs were implicitly generated.[71] The two multisequence alignments consisted of 3776 and 12113 sequences for class I and class II, respectively, including 236 and 217 fusion proteins from class VII (also known as class IV) and VIII (also known as class III), respectively. As a validation of the accuracy of the standard numbering scheme, the assignment of standard numbers to positions of proteins with known crystal structure which were not part of the reference alignment was compared to the results from a structure alignment created by STAMP.[94] Thus, pairwise alignments of 26 class I and 14 class II CYPs to their respective reference sequence were performed using the numbering scheme and STAMP. Subsequently, the results from both methods were compared by checking the sequence identity for each column of the resulting alignments. The average difference of the alignment columns between the two alignment methods was used as a validation criterion allowing for a position specific evaluation of the alignment quality. The validation for class I CYPs was performed starting from position 10, because structure information on the N-terminal residues was lacking in the PDB entry 1PHG of CYP101A1. For class II CYPs, the validation was performed starting from position 2 (Figure S27 on pages 149 and 150).

Comparison between the class I and the class II numbering scheme

Separate standard numbering schemes for the two classes were developed for a better accuracy in variable regions and a more reliable annotation of structural elements. To relate the two numbering schemes, the structures of CYP101A1 and CYP102A1 (reference sequences for the two numbering schemes) were aligned by STAMP,[94] and all structurally equivalent positions were identified (Table S1 on pages 143-145). The equivalence table

shows structurally corresponding positions of the reference sequences and allows for comparison of sequences numbered using the two numbering schemes.

Conservation analysis

The CYPED was systematically analyzed for conserved positions based on multisequence alignments derived from the standard numbering scheme. The multisequence alignments of CYPs from class I (including fusion constructs from class VII) and class II (including fusion constructs from class VIII) were screened for columns where a single amino acid was highly conserved (> 80 %), or where amino acids with similar physicochemical properties were found in more than 80 % of all sequences. Four groups of different physicochemical properties were distinguished: hydrogen bonding residues (cysteine, asparagine, glutamine, serine, threonine, tyrosine), charged residues (aspartic acid, glutamic acid, histidine, lysine, arginine), aromatic residues (phenylalanine, tryptophan, tyrosine), and hydrophobic residues (alanine, glycine, leucine, isoleucine, methionine, valine).

5.1.4 Results

Update of the CYPED and family classification of cytochrome P450 monooxygenases

The Cytochrome P450 Engineering Database (CYPED) was updated using one seed sequence for each homologous family from version 2.03 of the CYPED.[72] The update resulted in 19198 sequences, which is an increase of 41 % as compared to the previous version. After removal of partial sequences, the updated CYPED consists of 16732 sequences, as compared to 11195 sequences in the previous version. The number of superfamilies and homologous families increased from 249 to 258 and from 620 to 626, respectively. The number of proteins with structure information increased from 50 (corresponding to 125 PDB entries) to 72 (corresponding to 408 PDB entries). From this dataset, family-specific multisequence alignments and HMM profiles were generated.

The numbering schemes for class I and class II of cytochrome P450 monooxygenases

All sequences in the CYPED were assigned to a class, based on a previously proposed classification scheme.[26] For the two largest classes, numbering schemes were developed: a class I-specific numbering scheme for class I and the fusion proteins in class VII (23% of all sequences), due to their high structural and functional similarity, and a class II-specific numbering scheme for class II and the fusion proteins in class VIII (72% of all sequences).[163]

The two numbering schemes were validated by comparison of the numbering scheme-based alignment to a structural alignment for a selected set of proteins with known structure (Figure S27 on pages 149 and 150). Despite the high sequence diversity, the validation revealed high similarity between the two alignment methods: in class I and class II, 86% and 81%, respectively, of the columns were identically aligned. The reliability of the alignments was lower near the N-terminus of the CYPs because of higher sequence diversity in this region and residues that are lacking in the respective PDB entry. In addition, there were several short regions where the alignments deviated, mostly loops that differ in length, sequence, and structure, and thus cannot always be superimposed adequately: positions 86-98 (αB' between the β1_5 and αC) and 187-191 (loop between αF and αG) in class I CYPs, and positions 70-84 (αB' between the β1_5 and αC), 165-171 (loop between αE and αF), 187-201 (loop between αF and αG), 226-231 (loop between αG and αH), and 382-389 (loop between meander loop and cysteine pocket) in class II.

Analysis of conserved positions in class I and class II CYPs

A conservation analysis was performed for the two classes of CYPs. The results were grouped into five types of positions (Table 13, 14, 15, 16, 17 on pages 81, 83, 84, 85 positions where one single amino acid was highly conserved in more than 80 % of all sequences, and four types of positions where amino acids with the same physicochemical property (aromatic, hydrogen-bonding, charged, hydrophobic) were conserved in more than 80 % of all sequences. Conserved positions were found only on the proximal surface and in the core of the CYPs, but not on the distal surface of class I and class II CYPs (Figure 18 on the following page).

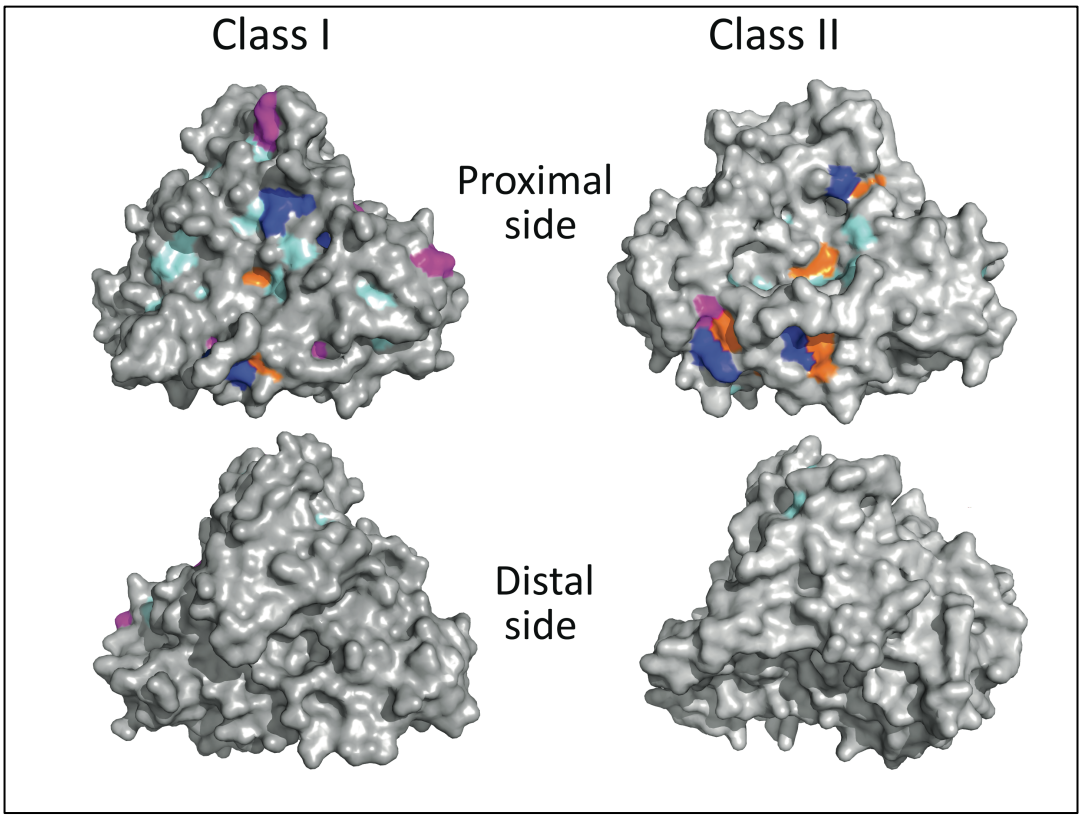


Figure 18: Conservation on proximal and distal surface of class I and class II CYPs. Amino acids conserved in class I were highlighted on the structure of CYP101A1 (PDB code 1PHG) and in class II on the structure of CYP102A1 (PDB code 1ZOA). Amino acids are colored by properties magenta: negative charge, blue: positive charge, orange: aromatic, pale blue: hydrophobic.

Highly conserved positions

In class I and II, respectively 16 and 17 positions were found with a single amino acid conserved in more than 80 % of all sequences. Among these positions, 12 were equivalent in class I and class II, 4 were class I-specific, and 5 were class II-specific (Table 13 on page 81).

Standard position H108 was conserved in 84 % of class I CYPs and was described to interact with heme.[38] This position corresponds to class II position W96 (conserved in 80 % of all class II CYPs) which was described to have a role in the incorporation of heme.[76] Position R112 in class I CYPs (94 % conserved) corresponding to R100 in class II CYPs (84 %) was described to interact with heme (Figure 14) and to be essential for high electron transfer rates in CYP101A1.[38,168] E287 (98 %) in class I and the corresponding E320 (99 %) in class II, as well as R290 (97 %) in class I and R323 (99 %) in class II form the ExxR motif.[158]

Standard position R342 (96 %) in class I CYPs was described to stabilize the protein structure[38] and corresponds to R378 (86 %) in class II CYPs. In both classes, the heme ligand

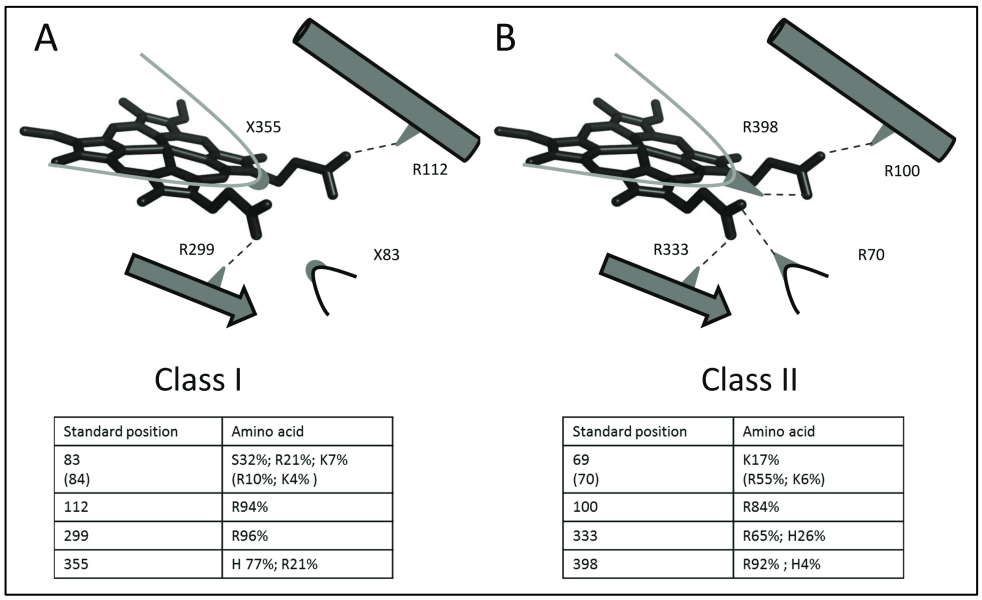
cysteine[169] is conserved: C357 (98 %) and C400 (99 %) in class I and class II CYPs, respectively. The results show that the ExxR motif and the heme ligand cysteine are not conserved in 100 % of the studied sequences due to misalignment, or because these highly conserved residues are missing in some members of the CYP family.[170–172] Furthermore, a threonine residue was highly conserved in class I (T252 in 85 % of all class I sequences) and in class II CYPs (T268 in 87 % of all class II sequences), as well as a phenylalanine residue in class I (F350 in 93 %) and in class II CYPs (F393 in 98 %), which were described as important for the catalytic machinery.[35,36] The conserved threonine was described to have multiple roles including substrate recognition,[34,173] proton donation,[174] and oxygen activation.[175] The conserved phenylalanine was described to interact with the heme-ligand cysteine and to modulate the reduction potential of heme.[176,177] In addition, there were three highly conserved glycines and a highly conserved proline at equivalent positions in class I and class II CYPs: G249 (82 %), P268 (85 %), G353 (98%), and G359 (96 %) in class I CYPs, corresponding to G265 (86 %), P284 (84 %), G396 (99 %), and G402 (94 %) in class II CYPs. Proline residues are known to disrupt α-helices and β-strands and mediate folding,[178,179] while glycine residues disrupt helices and contribute to sharp loops.[180] The first conserved glycine is located in the center of αI and can disturb the helix. The conserved proline is located at the N-cap of αJ. The second and third glycine residues are located in the cysteine pocket (Table 18 on page 88) and contribute to the characteristic conformation of this loop.

Position R299 was highly conserved in class I CYPs (96 %) and was described to stabilize heme (Figure 19 on page 82).[38] In class II CYPs, the corresponding position 333 was less conserved (R 65 %, H 26 %, K 3 %), which indicates that arginine is essential for class I CYPs but can be replaced by histidine and lysine in class II CYPs. A363 was highly conserved in class I CYPs (85 %) and was shown to have an important role in function and stability of CYP101A1.[65] It corresponds to the less conserved position A406 in class II CYPs (74 %). There are also two conserved glycine residues that are present in class I CYPs, G315 (86 %) and G351 (92 %). Those positions correspond to moderately conserved glycines in class II CYPs, G350 (76 %) and G394 (64 %). The first conserved glycine residue is located at the N-cap of β1_3. The second conserved glycine is located in the cysteine pocket (Table 13 on the following page).

Table 13: Positions with single amino acids conserved in over 80% of sequences. Standard positions structurally corresponding between the classes are in the same rows. Positions with single amino acids conserved in over 80% of sequences are in bold, the corresponding less conserved positions are in *italic*.

Class I		Class II		Function
Standard position	Amino acid	Standard position	Amino Acid	
108	**H 84%**	96	**F 80%**	Heme interaction/incorporation [76,181]
112	**R 94%**	100	**R 84%**	Heme propionate interaction, electron transfer [168,181]
249	**G 81%**	265	**G 85%**	Structurally significant residue
252	**T 85%**	268	**T 87%**	Active site threonine [34,35,174,182]
268	**P 85%**	284	**P 84%**	Structurally significant residue
277	*D 36% E 33%*	293	**E 95%**	Function not described in literature
287	**E 98%**	320	**E 99%**	ExxR motif [39]
290	**R 97%**	323	**R 99%**	ExxR motif [39]
299	**R 96%**	*333*	*R 65%*	Heme propionate interaction [38]
315	**G 86%**	*350*	*G 74%*	Structurally significant residue
338	*F 67%*	374	**F 94%**	Function not described in literature
340	*P 37%*	376	**P 95%**	Structurally significant residue
342	**R 96%**	378	**R 86%**	Structure stabilization [38]
349	*P 25%*	392	**P 82%**	Structurally significant residue
350	**F 93%**	393	**F 98%**	Heme-ligand cysteine interaction [177,183]
351	**G 92%**	*394*	*G 61%*	Structurally significant residue
353	**G 98%**	396	**G 99%**	Structurally significant residue
355	*H 75% R 22%*	398	**R 92%**	Heme propionate interaction [18,73]
357	**C 98%**	400	**C 99%**	Heme ligand [169]
359	**G 96%**	402	**G 94%**	Structurally significant residue
363	**A 85%**	*406*	*A 72%*	Stability of CYP101 [65]

R398 was highly conserved in class II CYPs (92 %). It is close to the heme and was described to stabilize the heme in substrate-free CYP102A1 (Figure 19 on the following page),[18,73] whereas at the corresponding position 355 in class I CYPs histidine and arginine were found in 77 % and 20 % of the sequences, respectively, suggesting that arginine is crucial for class II CYPs but not for class I CYPs. Furthermore there are two class II-specific prolines in positions P376 (96 %) and P392 (82 %), which correspond to the less conserved positions 340 (36 %) and 349 (28 %) in class I CYPs. The first proline is located in the meander loop, the second in the cysteine pocket (Table 16 on page 84). Additionally, two conserved positions that have not yet been described in the literature were found to be conserved in class II CYPs, E293 (95 %) and F374 (94 %). These residues are located on αJ and in the meander loop. In class I CYPs, the corresponding positions are less conserved (Table 13) suggesting that these residues are essential for class II CYPs, but not for class I CYPs.



Standard position	Amino acid
83 (84)	S32%; R21%; K7% (R10%; K4%)
112	R94%
299	R96%
355	H 77%; R21%

Standard position	Amino acid
69 (70)	K17% (R55%; K6%)
100	R84%
333	R65%; H26%
398	R92% ; H4%

Figure 19: Schematic representation of conserved arginine residues in the heme environment in A) class I and B) class II CYPs. Heme interacting residues corresponding between the classes are in the same rows of the tables. In class II CYPs, heme propionate A can be involved in interaction with lysine at position 69 or with arginine at position 70. The amino acid distributions at position 70 in class II and the corresponding position 84 in class I are shown in parenthesis.

Aromatic residues

There was only one position with an aromatic residue conserved in class I and in class II CYPs (Table 14 on the following): position 332 in class I CYPs (F 60 %, W 12 %, Y 13 %) corresponding to position 367 in class II CYPs (W 56 %, F 27 %, Y 15 %). This position is localized on the meander loop, but its function has not yet been described in literature. In addition, there were four positions in class II CYPs with conserved aromatic amino acids: position 275 (W 51 %, F 22 %, Y 8 %), 313 (Y 77 %, F 9 %), 379 (F 78 %, W 17 %, Y 2 %), and 421 (F 73 %, Y 16 %). These positions correspond to variable positions in class I CYPs, and their functions has not yet been described in literature. Thus, aromatic residues are considerably more conserved in class II CYPs than in class I CYPs.

Table 14: Conservation of aromatic amino acids. Corresponding positions are in the same rows, conserved positions are in bold. Empty fields indicate amino acid frequencies between 0 and 1 %.

Class I				Class II				Function
Standard position	F	W	Y	Standard position	F	W	Y	
108		**13%**		**96**	**2%**	**80%**	**1%**	Heme interaction/incorporation[38,76]
259	*6%*	*19%*	*1%*	**275**	**22%**	**51%**	**8%**	Function not described in literature
280	*3%*		*11%*	**313**	**9%**		**77%**	Function not described in literature
332	**60%**	**12%**	**13%**	**367**	**27%**	**55%**	**15%**	Function not described in literature
338	*67%*		*6%*	**374**	**94%**		**5%**	Function not described in literature
---	---	---	---	**379**	**78%**	**17%**	**2%**	Function not described in literature
350	**93%**		**2%**	**393**	**98%**	**1%**	**1%**	Heme-ligand cysteine interaction[177,183]
378	*32%*	*1%*	*10%*	**421**	**72%**		**16%**	Function not described in literature

Hydrogen-bonding amino acids

Only one position with a conserved hydrogen-bonding residue was found in class I and class II CYPs (Table 15). Position 253 in class I CYPs (T 78 %, S 5 %) corresponding to position 269 in class II CYPs (T 57 %, S 23 %) is located close to other catalytically significant positions and has not yet been described in literature.

Table 15: Conservation of hydrogen bonding amino acids. Corresponding positions are in the same rows, highly conserved positions are in bold. Empty fields indicate amino acid frequencies of less than 1 %.

Class I						Class II						Function
Standard position	C	N	Q	S	T	Standard position	C	N	Q	S	T	
252	**---**	**1%**	**1%**	**3%**	**85%**	**268**	**---**	**3%**	**---**	**5%**	**88%**	Active site threonine [35,174,182]
253	**---**	**---**	**---**	**5%**	**78%**	**269**	**---**	**1%**	**---**	**23%**	**57%**	Function not described in the literature
357	**98%**	**---**	**---**	**---**	**---**	**400**	**99%**	**---**	**---**	**---**	**---**	Heme ligand [169]

Charged amino acids

In addition to the highly conserved charged positions, there were also positions where multiple charged amino acids are conserved in over 80% of all sequences (Table 16 on the following page). Only one position with a charged residues was conserved in class I and class II CYPs. Position 251 in class I CYPs (E 56,7 %, D 33,0 %) corresponding to position 267 in class II CYPs (D 46 %, E 43 %, H 5 %), which plays a role in the stabilization of water and is part of the proton delivery system.[18] Four positions with negatively charged amino acids were

conserved in class I CYPs: 104 (D 68 %, E 15 %), 227 (E 37 % D 27 %, R 8 %, K 4 %, H 5 %), 304 (D 73 %, E 10 %), and 328 (D 79 %, R 4 %, H 3 %, E 2 %).

Table 16: Conservation of charged amino acids. Corresponding positions are in the same row, highly conserved positions are in bold. Empty fields indicate amino acid frequencies of less than 1 %.

Class I						Class II						Function
Standard position	D	E	H	K	R	Standard position	D	E	H	K	R	
104	**68%**	**16%**	**---**	**1%**	**---**	90	15%	6%	4%	3%	6%	Function not described in the literature
108	**---**	**---**	**85%**	**---**	**---**	96	**---**	**---**	8%	**---**	**---**	Heme interaction/incorporation [38,76]
112	**---**	**---**	**---**	**4%**	**94%**	**100**	**---**	**---**	**6%**	**5%**	**84%**	Heme propionate interaction, electron transfer [38,168]
227	**27%**	**37%**	**5%**	**4%**	**8%**	241	2%	7%	1%	15%	10%	Function not described in the literature
251	**33%**	**57%**	**1%**	**---**	**---**	**267**	**46%**	**43%**	**2%**	**---**	**---**	Stabilization of water in proton delivery [18]
273	1%	8%	1%	13%	24%	**289**	**---**	**3%**	**---**	**57%**	**24%**	Function not described in the literature
277	36%	33%	2%	2%	5%	**293**	**1%**	**95%**	**---**	**---**	**---**	Function not described in the literature
287	**---**	**98%**	**---**	**---**	**---**	**320**	**---**	**99%**	**---**	**---**	**---**	ExxR motif [39]
290	**---**	**---**	**---**	**---**	**97%**	**323**	**---**	**---**	**---**	**---**	**99%**	ExxR motif [39]
299	**---**	**---**	**---**	**---**	**96%**	**333**	**---**	**---**	**26%**	**3%**	**65%**	Heme interaction [38]
304	**73%**	**10%**	**---**	**---**	**1%**	338	63%	6%	1%	1%	1%	Function not described in the literature
328	**79%**	**1%**	**3%**	**---**	**4%**	363	64%	1%	**---**	**---**	2%	Function not described in the literature
342	**---**	**---**	**---**	**---**	**96%**	**378**	**---**	**---**	**10%**	**---**	**86%**	Structure stabilization [181]
355	**---**	**---**	**77%**	**---**	**21%**	**398**	**---**	**---**	**4%**	**---**	**92%**	Heme propionate interaction [18,73]

In class II CYPs, the positions corresponding to 304 and 328 were negatively charged (position 338: 69 %, position 363: 65 %), while the positions corresponding to 104 and 227 were negatively charged in less than 20 % of class II CYPs. Position 289 (K57 %, R 24 %, E 3 %) was the only class II-specific position with a charged amino acid. The corresponding position in class I was variable and the function of these positions has not yet been described in literature. All charged positions are located on the protein surface.

Hydrophobic amino acids

The conserved hydrophobic residues mostly play a role in structure stabilization, since most of them were located in the protein core. Conserved hydrophobic amino acids were also found in the cysteine pocket and possibly contribute to heme stabilization. The surface-exposed hydrophobic residues might also play a role in the CYP-redox partner interactions. In class I CYPs, 53 conserved positions with hydrophobic amino acids were found, in class II only 33 (Table 17 on the following page). 25 of these positions were corresponding between class I and class II.

Table 17: Conservation of hydrophobic amino acids. Corresponding positions are in the same rows, highly conserved positions are in bold. Empty fields indicate amino acid frequencies of less than 1 %.

Class I							Class II						
Standard position	A	G	I	L	M	V	Standard position	A	G	I	L	M	V
52	**4%**	**---**	**18%**	**18%**	**2%**	**46%**	*38*	*2%*	*3%*	**---**	**---**	**---**	**---**
63	*2%*	**---**	*4%*	*3%*	**---**	*2%*	**50**	**3%**	**---**	**16%**	**20%**	**1%**	**44%**
65	**9%**	**---**	**14%**	**20%**	**2%**	**45%**	**52**	**6%**	**1%**	**22%**	**22%**	**1%**	**29%**
71	**18%**	**1%**	**22%**	**8%**	**1%**	**37%**	**58**	**28%**	**---**	**24%**	**4%**	**3%**	**24%**
74	**13%**	**2%**	**13%**	**25%**	**2%**	**42%**	**61**	**15%**	**---**	**26%**	**7%**	**3%**	**33%**
100	**---**	**---**	*14%*	*37%*	*15%*	*14%*	**86**	**---**	**---**	**27%**	**39%**	**7%**	**14%**
114	**13%**	**---**	**13%**	**44%**	**8%**	**7%**	*102*	*5%*	**---**	*17%*	*24%*	*8%*	*8%*
115	**4%**	**2%**	**8%**	**33%**	**8%**	**37%**	*103*	*14%*	**---**	*8%*	*29%*	*4%*	*9%*
124	**4%**	**---**	**20%**	**24%**	**12%**	**31%**	**112**	**1%**	**1%**	**10%**	**43%**	**16%**	**16%**
131	**3%**	**---**	**44%**	**9%**	**7%**	**27%**	*119*	*1%*	**---**	*25%*	*5%*	*12%*	*15%*
139	**3%**	**---**	**23%**	**37%**	**3%**	**26%**	*127*	*3%*	**---**	*16%*	*23%*	*6%*	*28%*
162	**9%**	**2%**	**13%**	**4%**	**6%**	**56%**	*152*	*4%*	**---**	*25%*	*6%*	*4%*	*33%*
163	**10%**	**---**	**59%**	**13%**	**2%**	**7%**	**153**	**6%**	**2%**	**58%**	**10%**	**5%**	**9%**
166	**8%**	**---**	**8%**	**45%**	**15%**	**12%**	*157*	*29%*	*3%*	*10%*	*13%*	*5%*	*15%*
167	**5%**	**---**	**6%**	**57%**	**11%**	**5%**	*158*	*3%*		*2%*	*14%*	*8%*	*3%*
169	**4%**	**12%**	**13%**	**13%**	**3%**	**41%**	*159*	*1%*	*73%*	**---**	**---**	**---**	**---**
208	**7%**	**---**	**40%**	**12%**	**2%**	**27%**	**220**	**2%**	**---**	**47%**	**11%**	**4%**	**26%**
219	**6%**	**---**	**14%**	**52%**	**6%**	**12%**	*233*	*3%*	**---**	*6%*	*27%*	*10%*	*3%*
220	**6%**	**2%**	**19%**	**41%**	**7%**	**8%**	**234**	**2%**	**---**	**23%**	**42%**	**7%**	**14%**
223	**1%**	**2%**	**10%**	**70%**	**10%**	**2%**	*237*	**---**	**---**	*6%*	*50%*	*11%*	*3%*
224	**14%**	**1%**	**26%**	**20%**	**3%**	**19%**	**238**	**1%**	**---**	**18%**	**62%**	**10%**	**3%**
233	**1%**	**---**	**9%**	**61%**	**15%**	**5%**	*249*	*1%*	**---**	*9%*	*40%*	*6%*	*3%*
238	**12%**	**---**	**29%**	**36%**	**2%**	**13%**	**254**	**2%**	**---**	**43%**	**32%**	**6%**	**13%**
245	**2%**	**2%**	**9%**	**64%**	**9%**	**9%**	*261*	**---**	**---**	*13%*	*31%*	*13%*	*5%*
246	**4%**	**2%**	**15%**	**53%**	**7%**	**8%**	*262*	**---**	**---**	*13%*	*23%*	*12%*	*5%*
248	**75%**	**18%**	**---**	**---**	**---**	**---**	**264**	**74%**	**14%**	**---**	**---**	**---**	**1%**
249	**12%**	**82%**	**1%**	**---**	**---**	**---**	**265**	**9%**	**86%**	**---**	**---**	**---**	**---**
257	**2%**	**---**	**39%**	**39%**	**8%**	**3%**	**273**	**3%**	**---**	**15%**	**52%**	**11%**	**8%**
261	*9%*	*1%*	*7%*	*31%*	*8%*	*22%*	**277**	**2%**	**---**	**11%**	**43%**	**18%**	**7%**
264	**1%**	**2%**	**5%**	**78%**	**7%**	**2%**	**280**	**---**	**---**	**5%**	**76%**	**14%**	**3%**
265	*23%*	*2%*	*3%*	*36%*	*1%*	*3%*	**281**	**30%**	**5%**	**9%**	**21%**	**8%**	**9%**
274	**4%**	**---**	**5%**	**69%**	**1%**	**15%**	**290**	**24%**	**---**	**10%**	**33%**	**1%**	**25%**
278	*7%*	*9%*	**---**	**---**	**---**	**---**	**294**	**2%**	**---**	**41%**	**22%**	**3%**	**25%**
---	**---**	**---**	**---**	**---**	**---**	**---**	**311**	**1%**	**---**	**1%**	**56%**	**37%**	**1%**
281	**9%**	**---**	**25%**	**28%**	**5%**	**14%**	*314*	**---**	**---**	*4%*	*58%*	*7%*	*6%*
359	**2%**	**96%**	**---**	**---**	**---**	**---**	*403*	*4%*	**---**	*6%*	*5%*	*12%*	*3%*
362	**---**	**---**	**9%**	**69%**	**1%**	**5%**	**406**	**74%**	**19%**	**---**	**---**	**---**	**2%**
363	**85%**	**6%**	**---**	**---**	**---**	**1%**	*407*	*1%*	**---**	*7%*	*23%*	*14%*	*4%*
365	**8%**	**---**	**4%**	**39%**	**20%**	**9%**	*408*	*12%*	**---**	*4%*	*8%*	*21%*	*7%*
367	**20%**	**7%**	**10%**	**27%**	**21%**	**6%**	**410**	**12%**	**---**	**11%**	**28%**	**19%**	**17%**
369	**7%**	**---**	**31%**	**15%**	**2%**	**24%**	*412*	*5%*	**---**	*12%*	*37%*	*4%*	*17%*

Class I							Class II						
Standard position	A	G	I	L	M	V	Standard position	A	G	I	L	M	V
370	**38%**	**7%**	**6%**	**13%**	**4%**	**16%**	*413*	*12%*	*7%*	*11%*	*10%*	*6%*	*14%*
371	*2%*	**---**	*7%*	*57%*	*2%*	*6%*	**414**	**4%**	**---**	**11%**	**53%**	**5%**	**11%**
374	**---**	**---**	**13%**	**60%**	**5%**	**9%**	**417**	**---**	**---**	**21%**	**51%**	**6%**	**9%**
375	6%	**---**	6%	49%	1%	7%	**418**	**3%**	**---**	**10%**	**61%**	**3%**	**18%**

5.1.5 Discussion

Class assignment

There are ten classes of CYPs described in the literature[8] and this classification was used as the base for the class assignment in this study. Since not all CYPs have been characterized experimentally and the respective redox partners are not known for all the sequences in the CYPED a coherent class assignment strategy combining available literature data, the information about the source organism and the sequence similarity to characterized CYPs was applied. It was previously reported that the length of the αJ' and the meander loop could be the base for the class assignment[19], but some CYPs (i.e. class I eukaryotic mitochondrial CYPs) exhibit lengths of those elements which do not correlate with the redox partner type. Due to this observation, the approach using the sequence length of defined regions for class assignment was not implemented. It is possible that our comprehensive approach does not include all non-typical CYPs but up to now only 753 out of 16732 (<5%) of the sequences in CYPED were assigned to non-typical classes (not class I, II or fusion).

Conservation analysis

The conservation analysis confirmed well-documented functionally relevant positions such as the heme ligand cysteine[169] or the ExxR motif,[158] but also revealed positions which are not yet known as conserved nor have been investigated by mutational studies. All conserved positions are localized on the proximal surface and in the core of class I and class II CYPs (Figure 18 on page 79), while the distal surface involved in substrate recognition is variable. This is in accordance with the fact that all functions that are common to all CYPs are located in the core and at the proximal side (heme binding, catalytic machinery, redox partner interactions), while CYPs differ widely in their substrate spectrum. Despite the high sequence diversity inside the CYP classes, there were several class-specific positions that were

conserved in more than 80% of the sequences of one class, but more variable in the other class.

Conserved residues on the proximal surface of CYPs

The heme domain of all class I CYPs accepts two electrons from an iron-sulfur-cluster protein, class II CYPs from a CPR-type reductase. There has been much interest in identifying proximal surface residues that influence the activity and electron transfer rates of CYPs, and it has been pointed out that in both classes mostly positively charged residues on the proximal CYPs surface are involved in the interaction with the electron donor.[93,98,184,185,69] We expected to find conserved class-specific positively charged residues on the proximal surface of CYPs because the iron-sulfur cluster redox partner differs considerably from the CPR-type reductase, and because in most organisms multiple CYPs of the same class interact with a single reductase.[186,187] Although we found two class I-specific (positions 227 and 304) and three class II-specific (positions 289, 293, 313) conserved residues on the proximal surface of the heme domain, none of them has yet been described in literature to be involved in the interaction between heme domain and redox partner, and only one class II-specific residue is positively charged.

The two class I-specific conserved residues on the proximal surface are not near to each other nor in the vicinity of the site where heme is close to the protein surface (C357 in class I CYP101A1) and where a close contact to the redox partner is expected to occur for proper electron transfer[38] (20,6A for position 227 and 28,5A for position 304). Since the iron-sulfur cluster protein is a small (*P. putida* ferredoxin is 107 amino acids), it would not be possible for the protein to be properly arranged relatively to position 357 and interact with positions 227 and 304.[188]

The class II-specific proximal surface residues are in close proximity to each other and form a patch of three amino acids that has not yet been described in literature. The patch consists of: a positively charged lysine or arginine in position 289, a negatively charged glutamic acid in position 293, and an aromatic residue in position 313. The conserved residues in the patch do not interact with each other but could be involved in the interaction with the redox partner. Additionally, these three residues are accompanied by three class-specific hydrophobic non-surface residues in positions 290, 294 and 311. Since CPR is a large, two-domain protein, it is possible that this patch, which is not in direct proximity to the heme, might be involved in the interaction with CPR. The function of those residues has not yet been studied, so it would be

extremely interesting to study the effect of mutations in those positions on the interaction with redox partner.

The small number of class-specific conserved positions on the proximal surface indicates that redox partner recognition is not highly conserved inside class I or class II CYPs, but there are CYP-specific features such as redox potential[74] or loop lengths[72] that can mediate or prevent productive interactions between heme domain and redox partner.

Cysteine pocket

The cysteine pocket is characteristic structural element of all CYPs, it is a pocket shaped structural element with heme binding cysteine. Proline and glycine residues are often found in short loops and in the cap regions of α–helices and β–strands.[178–180] The cysteine pocket is rich in glycine and proline residues, and the conservation patterns of those residues in the cysteine pocket differ between class I and class II CYPs (Table 18). Since proline residues significantly influence the shape of the protein backbone and glycine residues influence its flexibility, we propose that the observed differences might influence flexibility and shape of the cysteine pocket. It would be interesting whether there are optimal proline-glycine motifs for class I or class II CYPs.

Table 18: Class specific conservation of glycine and proline residues in the cysteine pocket of CYPs. Structurally corresponding residues are in the same columns, differences between classes are highlighted grey.

Class I	349 P25%	350 F93%	351 G92%	…	353 G98%	…	357 C98%	…	359 G96%
Class II	392 P82%	393 F98%	394 G61%	…	396 G99%	…	400 C99%	…	402 G94%

Structurally relevant residues

The conserved hydrophobic residues contribute to the hydrophobic core of the protein, thus stabilizing the structural elements, α-helices and β-strands. Besides the hydrophobic amino acids there are only two structurally relevant positions that are conserved in both classes. The first position is occupied by a conserved aromatic residue (class I: position 332, class II: position 367), the second position by a conserved arginine residue (class I: standard position 342, class II: standard position: 378). The arginine residue was reported to stabilize the CYP structure by interaction with the conserved glutamic acid of the ExxR motif.[158] These two residues are involved in a complex cluster of conserved amino acids in class II proteins,

where besides the aromatic residue conserved in both classes two additional conserved positions with aromatic amino acids were found (in positions 374 and 379). These aromatic residues have not yet been described in literature. Their role might be the stabilization of the structure between the conserved arginine and the glutamic acid. The difference to class I CYPs where the additional aromatic residues are less conserved might be associated with a different reductase interaction because class II CYPs interact with the much bigger CPR reductase as compared to the small iron-sulfur-cluster protein and better structure stabilization might be needed in order to facilitate the interaction.

There were also structurally relevant class I-specific residues: a conserved glycine at the N-cap of β1_3 (position 315) and two negatively charged residues (position 104 and 328) in loops close to the proximal surface. The conserved glycine residue at position 315 allows for a sharp turn near to β1_3. Interestingly, the corresponding position in class II is less conserved. The function of two negatively charged residues has not yet been described in literature. From structure of CYP101A1[38] we predict that these residues might be involved in salt bridges: position 328 interacts with arginine at position 67 (positively charged residues in more than 45% of sequences), position 104 might interact with serine at position 82 or 83 (49 and 33%, respectively). To our knowledge, those residues have not yet been studied in CYP101A1, so it would be interesting to probe their role in structure stabilization. Two additional class II-specific conserved aromatic residues in the protein core (position 275 and 421) could also be involved in structure stabilization, as described for many proteins.[75,189]

Heme-interacting residues

Most of the residues that interact with the catalytically active heme or are part of the catalytic machinery of CYPs were highly conserved in all CYPs. Four highly conserved residues were described in literature that are located close to the heme and are part of the catalytic machinery: the negatively charged amino acid involved in the proton delivery (class I position 251, class II position 267),[18] the conserved threonine involved in substrate recognition, proton donation, and oxygen activation (class I position 252, class II position 268),[34,35,173–175] the conserved phenylalanine interacting with the heme ligand cysteine (class I position 350, class II position 393),[35,176,177] and the heme ligand cysteine (class I position 357, class II position 400).[169] In addition, there is one highly conserved residue that has not been described yet, a hydrogen bonding residue (class I position 253, class II position 269) which is located in close proximity to the well-described conserved hydrogen bonding residue (class I position 252,

class II position 268) and the conserved negatively charged residue (class I position 251, class II position 267), which both are involved in proton delivery.[18,34–36,173–175] Since the side chain of this residue is pointing towards the heme, we propose that it also might be part of the catalytic machinery.

Despite the high conservation of the catalytic machinery in all CYPs, there were significant class-specific differences. Most remarkably, the heme-interacting histidine at position 108[38] is conserved in class I CYPs and corresponds to a conserved tryptophan at position 96 in class II CYPs which was suggested to interact with heme and to be involved in heme incorporation.[76] Therefore, it would be interesting to assess the effect of mutations at these positions in class I or class II CYPs.

In addition, there were systematic differences in the four positively charged amino acids that interact with the two heme propionates (Figure 19 on page 82). The highly conserved arginine at position 112 in class I CYPs was described to interact with the propionate D moiety of heme and to have a significant influence on electron transfer rates[54]. The corresponding position 100 in class II CYPs, however, is less conserved. The functional relevance of a positive charge at this position was demonstrated for the class I CYP101A1, where a substitution of arginine to lysine resulted in an active CYP mutant, although with increased redox potential, while all other mutants were substantially less active[54]. In most of class II proteins there is additional residue interacting with the heme propionate D. This residue is a highly conserved arginine at class II position 398,[18,73] in class I this position is mostly occupied by histidine and in significantly lower degree arginine. CYPs from class I and class II also differ in the residues interacting with the heme propionate A. While in class I CYPs heme interacting arginine at position 299[38] is conserved in 96%, arginine or histidine occur in the corresponding position 333 in class II CYPs. In class II CYP102A1, a second residue (lysine in position 69) interacts with propionate A. However, this lysine was only found in 13% of class II CYPs. In most other class II CYPs such as CYP 2A6[55] the second propionate A – interacting residue is arginine or lysine at position 70. While positions 69 and 70 are conserved in class II CYPs, the corresponding positions 83 and 84 in class I CYPs are significantly less conserved. Because the heme-interacting residues influence the electron transfer rates[54], their class-specific conservation might be a direct consequence of the different electron donor proteins.

5.1.6 Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's 7th Framework Programme (FP7/2007-2013) under REA Grant Agreement 289217.

5.2 Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining

5.2.1 Abstract

Cytochrome P450 monooxygenases (CYPs) are a large, highly diverse protein family with a common fold. The sequences, structures, and functions of CYPs have been extensively studied resulting in more than 53000 scientific articles. A sequence-based literature mining algorithm was designed to systematically analyze this wealth of information on SNPs, designed mutations, structural interactions, or functional roles of individual residues. Structurally corresponding positions in different CYPs were compared and universal selectivity-determining positions were identified. Based on the *Cytochrome P450 Engineering Database (www.CYPED.BioCatNet.de)* and a standard numbering scheme for all CYPs, 4000 residues in 168 CYPs mentioned in 2400 articles could be assigned to 440 structurally corresponding standard positions of the CYP fold, covering 96% of all standard positions. 17 individual standard positions were mentioned in the context of more than 32 different CYPs. The majority of these most frequently mentioned positions are located on the six substrate recognition sites and are involved in control of selectivity, such as the well-studied position 87 in CYP102A1 ($P450_{BM-3}$) which was mentioned in the articles on 63 different CYPs. The recurrent citation of the 17 frequently mentioned positions for different CYPs suggests their universal functional relevance.

5.2.2 Introduction

Cytochrome P450 monooxygenases (CYPs) constitute a large enzyme family with high sequence diversity and high structural similarity. CYPs are found in all kingdoms of life and catalyze the oxidation of a broad spectrum of substrates.[192] The reactions catalyzed by CYPs are often highly regio- and stereoselective, therefore CYPs are of high interest for synthetic applications.[10] Their crucial role in xenobiotic metabolism makes them prominent enzymes in drug development.[153,154] Due to the synthetic potential of CYP-catalyzed reactions,

understanding the molecular basis of catalytic activity, substrate specificity, and regio- and stereoselectivity is of high interest.[193]

All CYPs share a common structural fold, which houses the catalytically active heme and a buried substrate binding pocket. The highly conserved CYP fold consists of thirteen α–helices and five β–sheets.[38,39] It allows for a sequence comparison of different family members despite their low sequence identity and for the unambiguous identification of six substrate recognition sites (SRSs) and cysteine pocket in all CYPs.[40,59] The SRSs were first introduced based on the experimental characterization of the superfamily CYP2 and the first crystal structure of CYP101A1.[38,40] SRS1 is located on the β-loop-α element formed by β-strand B1_5 and α-helix B' (naming of the secondary structure elements according to Poulos et al.[38]), SRS2 is located on α-helix F, SRS3 on α-helix G, SRS4 on α-helix I, SRS5 covers β-strand 1_4 and the neighboring loops, and SRS6 spans over β-strands 4_1 and 4_2. SRS2 and 3 are part of the entrance channel leading from the bulk to the binding pocket, whereas SRS1, 4, 5, and 6 form the access to the heme. The residues located in the SRSs interact directly with a ligand molecule, or determine the structure or flexibility of the binding site. The cysteine pocket is a structural element of the CYP fold. It consists of a loop around the heme ligand cysteine, which is located before the α-helix-L. This region contains numerous conserved residues, mostly involved in interactions with heme.[31]

Due to the high structural and functional similarity of CYPs, it can be expected that the structurally corresponding residues have a similar molecular role. This refers to conserved residues involved in the catalytic mechanism like the heme ligand cysteine, but it also applies to the residues of the substrate recognition sites SRS 1-6. Specific SRS positions that are mentioned to mediate selectivity in a wide range of different CYPs will be referred to as 'universal selectivity-determining positions'. The identification of universal selectivity-determining positions would accelerate the design of promising variants or variant libraries with diverse properties especially for proteins without known structure, where it is not possible to study the protein-ligand interactions. Previously, a residue in SRS5 (position five after the conserved EXXR motif) was predicted by systematic sequence analyses to mediate selectivity in different CYP superfamilies.[59] This prediction was validated by designing a CYP102A1 minimal mutant library with a broad diversity in selectivity.[58] In order to enable comparative analyses of different CYP sequences and to facilitate transfer of knowledge about the functional role of individual residues, two class specific standard numbering schemes for CYPs were introduced.[31] The numbering schemes were developed for class I and

class II CYPs, which differ in the redox partner type, but also show class-specific structural features like the length of the α–helix J/J' and the meander insertion.[26,31,194] Class I CYPs are mainly prokaryotic and accept electrons from ferredoxin and ferredoxin reductase. Class II CYPs are mostly eukaryotic and accept electrons from diflavin cytochrome P450 reductases (CPR). CYPs from other classes were numbered using the class I or class II numbering schemes based on similarities in the electron transfer chain composition.[26,31] Based on the structure information and guided by family-specific profiles, a standard numbering scheme allow for the identification of structurally corresponding positions with similar functional roles based only on the protein sequence.[71] As opposed to standard sequence alignments, the numbering schemes allow for high quality alignments and position numbering within structurally conserved protein families even for representatives with low sequence identity,[31,71] thus allowing for identification of structurally corresponding positions.

The scientific literature was systematically analyzed to identify the most significant selectivity-determining positions in CYPs. This was based on the assumption that residues frequently mentioned in the literature for many different CYPs have a high probability of being functionally relevant. Thus, a sequence-based literature mining algorithm (SBLMA) was designed to extract literature information about residues and mutations mentioned for a wide range of CYPs. Similar literature mining approaches were previously implemented to extract disease-related mutations.[195–197] The extracted data were put into perspective of the whole family by analyzing the residues mentioned in the literature using the standard numbering schemes for CYPs,[31] thus identifying the most frequently mentioned and structurally corresponding positions in CYPs.

5.2.3 Methods

Update of the Cytochrome P450 Engineering Database

The Cytochrome P450 Engineering database (CYPED) was updated to provide resources for further analyses of the rapidly growing sequence space. Seed sequences representing different homologous families of CYPs were selected and subsequently used to acquire information on the entire sequence space of this vast enzyme family. Seed sequences were selected from searches in the NCBI GenBank[79] protein database using automatically generated queries for CYP names based on the Nelson nomenclature,[13] ranging from 'CYP1A1' through 'CYP1A5'

to ‘CYP5999Z5’. The resulting set of sequences was filtered based on sequence completeness criteria. Entries including the word ‘partial’ in their sequence description or with a sequence length of less than 300 residues were discarded. For each CYP name, a sequence compliant with these criteria was used as a seed sequence, resulting in 973 seed sequences. The database update was performed as described previously.[32] BLAST[198] searches for each of the seed sequences were performed against the non-redundant protein database of the NCBI GenBank[79] using an E-value cut-off of 10^{-5}. Entries with identical sequences were assigned to a single sequence entry. Sequence entries that were at least 98% identical were assigned to a single protein entry. Protein entries were named according to the name derived from the respective NCBI GenBank entries (e.g. "CYP1A1"). Short and long sequences (shorter than 300 amino acids and longer than 1300 amino acids) were deleted from the database. Protein structures from the Protein Data Bank (PDB)[199] were included into the database if the sequences were longer than 150 amino acids and shared at least 80% global sequence identity with any existing sequence in the database.

The data model of the CYPED consists of 4 hierarchical levels (sequences, proteins, homologous families, and superfamilies). The sequence identity cut-offs used for the family classification were based on the Nelson nomenclature.[13] The initial family classification was based on a clustering using *usearch.*[200] Homologous families were created by sequence clusters with at least 55% identity. Naming of the homologous families was mainly automated. In the first step, homologous families were named based on the respective protein entries (e.g. the homologous family "CYP1A" includes the proteins "CYP1A1" and "CYP1A2"; protein names such as "CYP1A1-like" were ignored). If a homologous family included proteins with conflicting names (such as "CYP1A1 and "CYP1B1") they were named in the subsequent steps. In the second step, yet unnamed homologous families were named based on proteins with crystal structure information. In the third step, the family classification was verified using the Needleman-Wunsch alignment method.[108] Calculation of the global identities of all sequence pairs in the database provided a comprehensive view of the family structure and allowed the merging of nearby homologous families. Homologous families were merged if their sequence identity was higher than 55% and grouped into the same superfamily if their sequence identity was higher than 40%. The superfamilies were named based on the respective homologous families (e.g. superfamily "CYP1" includes the homologous families "CYP1A" and "CYP1B"). Yet unnamed homologous families were named based on the most similar homologous family that was already named, unless sequence

identity was lower than 40%. In this case, it was labeled by adding the suffix ‘-like’ (e.g. "CYP1A-like"). If a homologous family was higher than 40% identical to homologous families with the ‘-like’ suffix, it was named by adding the ‘-like’ suffix to the respective superfamily (e.g. "CYP1-like").

The two class-specific standard numbering schemes were applied to all sequences in the database as described previously.[31] Accordingly, amino acid positions from different sequences (even from different classes of CYPs) could easily be compared according to their structural equivalence. The standard numbering schemes are based on the structure-guided sequence alignments and allow for identification of the structurally corresponding positions. Such positions are henceforth called ‘corresponding positions’. Moreover, the CYPED was implemented into the BioCatNet system, which allows for housing and analysis of family-specific protein databases.[78] The updated version is available online under the URL: www.CYPED.BioCatNet.de.

Sequence-based literature mining algorithm

To provide a link between the sequence data gathered in the CYPED and the abundance of literature about this protein family, a sequence-based literature mining algorithm (SBLMA) was designed (Figure 20 on page 100). The algorithm scans the available literature based on CYP names found in the CYPED, finds mentions of residues and mutations in the scientific literature, and matches the residues and articles with sequences in the database. The SBLMA can be divided into four main steps: (1) article search query generation, (2) acquisition of articles, (3) extraction of residues and mutations from articles, and (4) matching of the residues and articles to appropriate sequences.

In the first step of the algorithm, the descriptions of all CYP sequences found in the CYPED, as derived from the NCBI GenBank protein repository, are scanned to find possible CYP names. This scan is based on filters designed to find names according to the Nelson naming scheme[13] (e.g. "CYP101A1", "CYP6DJ1") or starting with a ‘P450’ prefix (e.g. "P450cam", "P450BM-3"). To minimize the number of false positive hits, preferentially full names based on the Nelson nomenclature were used (e.g. excluding "CYP101" or "CYP101A"). Additional names known from crystal structures were added to the list (e.g. "CYP119", "CYP125"). To further maximize search coverage, all entries in the query list were duplicated by replacing the respective ‘CYP’ or ‘P450’ prefix by its counterpart (e.g. "P450 102A1", "CYP BM-3"). This

part of the algorithm resulted in a list of CYPED-specific sequence identifiers and corresponding search queries.

The second step of the algorithm uses the list of search queries to acquire article abstracts and, if possible, full text and supplementary information. A list of PubMed identifiers (PMIDs) corresponding to the queries is acquired using the NCBI PubMed E-utilities application programing interface (API).[201] The result is a list of CYPED-specific sequence identifiers, search queries, and PMIDs. The PMIDs are used to acquire abstracts of the respective articles from PubMed. The abstracts are scanned for the presence of PubMed Central identifiers (PMCIDs) and digital object identifiers (DOIs). If a PMCID is found, the program tries to acquire the full text article from PubMed Central using the NCBI E-utilities API and supplementary materials using the Europe PubMed Central API. If it is not possible to acquire full text articles or no PMCID is available, the program tries to acquire the article using publisher-specific APIs. If such an API is available from the respective publisher, access to the full texts is attained either by use of PMID or DOI. In case where it is not possible to acquire a full article, only the abstract is used for the analysis.

The third step of the algorithm is designed to scan all available texts to find mentions of residues and mutations, and produce a list of positions mentioned in each article. During the scan, program sorts out strings that could be interpreted as amino acid positions but are false positives. This is based on a custom-designed dictionary containing vector names and cell line names. The main function of this step is based on a set of filters, which find mentions of amino acid positions and mutations in different formats (e.g. "F87A", "F87", "phenylalanine 87", "Phe87", "Phe at position 87"). To minimize the number of false positives, the search starts from position 20. In many CYPs N-termini contain transmembrane domains or long extended loops. Transmembrane domains might be functionally relevant,[202] but are not part of the CYP-fold and standard numbering schemes, and therefore do not allow for a straightforward identification of correspondence between positions from different sequences. At this stage the algorithm produced two lists, one containing the CYPED-specific sequence identifiers, search queries and PMIDs, and a second list containing PMIDs and putative mutations.

The fourth step of the algorithm matches found positions with sequence entries in the CYPED by checking if all amino acids mentioned at certain positions match the amino acids found in the respective sequence entry. After sorting out false positives based on the sequence

compatibility criterion, it is still necessary to manually check for false positives. The resulting manually curated list contains positions with corresponding sequence identifiers, CYP names, and PMIDs pointing to the articles. For analyses, the class-specific numbering schemes[31] were applied to all sequences with annotated positions.

5.2.2 Results

Identification of frequently mentioned positions by the sequence-based literature mining algorithm

The Cytochrome P450 Engineering Database was updated, resulting in 52674 sequences of 41513 proteins and 595 crystal structures. The sequence-based literature mining algorithm (SBLMA) (Figure 20 on page 100) was applied to integrate information about residues mentioned in the literature into the CYPED. In the first step of the SBLMA, more than 2000 unique CYP names corresponding to more than 7000 sequences were extracted from the CYPED. In the second step, more than 53000 PubMed articles corresponding to 66000 text files (abstracts, full text articles, and supplementary materials) were found to mention these CYP names in the title or abstract. In the third and fourth steps, mentioned residues and mutations were extracted from the respective text files and subsequently checked for compatibility with the sequences corresponding to the respective CYP names, resulting in 6000 residues mentioned in 4000 articles. Manual inspection further eliminated 30% of residues, which were false positives. The major source of false positives was due to the fact that the names of two CYPs with high sequence identity (such as "CYP11B1" and "CYP11B2") were referred to in the abstracts. Therefore, the residue mentioned in the article was compatible with both CYPs, but described for only one of them. Thus, the second CYP was identified as a false positive hit. The manual elimination of false positives finally resulted in 4000 residues from 168 CYPs, which were mentioned in 2400 articles (comprehensive list of all residues and the corresponding references is given in supplementary material Table S2[m]). In addition to the update, the CYPED was incorporated into the BioCatNet system.[78] To enable access to the results derived from the analyses done in this work, the web-accessible graphical user interface (*www.CYPED.BioCatNet.de*) was extended. For each sequence with residues identified in literature, a list of the respective positions and links to the publications are provided. Moreover, by making use of the standard numbering schemes, the web-interface allows for the transfer of available literature information from all sequences to the corresponding positions on the sequence of interest.

[m] Because of the length (over 150 pages) table and references are only available through the publishers' website.

SBLMA benchmark based on CYP102A1

To estimate the probability of false negatives, the SBLMA result for CYP102A1 (P450BM3) was compared to a literature review. The review provides a comprehensive list of mutations of 179 residues in CYP102A1 publications until mid of 2011.[80] 116 of these residues were also identified by the SBLMA (performed in November 2014), as well as an additional number of 82 residues which were not cited in the review. Combining both lists, we estimate that at least 261 residues have been mentioned in the literature up to now. 63 residues were not found by SBLMA, corresponding to at least 24% false negatives. This was caused by missing APIs to access full text articles and supplementary materials from some publishers.

CYP-fold coverage of the mentioned residues

The use of the automated SBLMA allowed us to assemble a comprehensive list of 4000 CYP residues mentioned in the literature (supplementary material Table S2). To put this information into perspective of the whole protein family, previously described class–specific standard numbering schemes for class I and class II CYPs were used.[31] This allowed identifying structurally corresponding positions among the residues mentioned in the articles about different CYPs. To allow for a comparison between class I and class II CYPs, a conversion table was established (Supplementary material Table S1 on pages 143-145). Based on a structural alignment between the two reference structures of the class I and class II numbering scheme, CYP101A1 and the heme domain of the CYP102A1 natural fusion protein, structurally corresponding positions were identified.[31] Since the class II numbering scheme includes longer loops than the class I scheme (458 versus 414 positions respectively), all class I standard positions were converted to class II standard positions which is equivalent to the CYP102A1 position numbering.

3638 residues mentioned in class I or class II CYPs were assigned to 440 standard positions that cover 96% of the CYP fold. Additional 326 residues mentioned in literature were not covered by the standard numbering schemes. 157 of those residues were located in loop regions, 90 residues in the N-terminal domain of CYPs, and 79 residues in the reductase domain of CYP102A1. Thus, more than 90% of all residues mentioned in the literature were successfully assigned to standard positions corresponding to positions in CYP102A1.

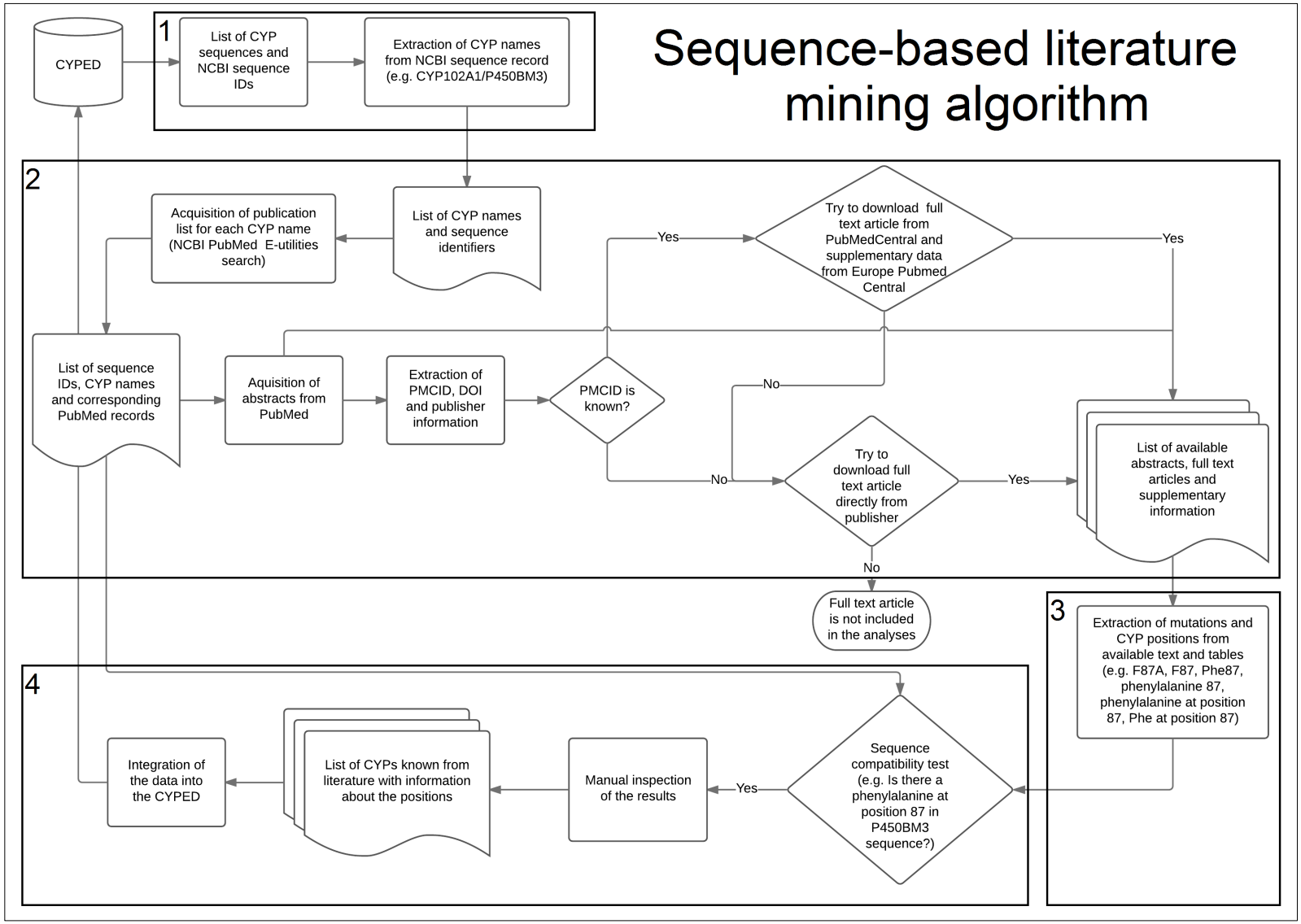


Figure 20: Flowchart of the sequence-based literature mining algorithm. The algorithm is divided into four main parts: (1) article search query generation, (2) acquisition of articles, (3) extraction of amino acid positions from articles and (4) matching of the positions and articles to appropriate sequences.

Most frequently mentioned positions in CYPs

The average number of different CYPs mentioned per position in the literature was the highest for the 98 positions in the substrate recognition sites and the 14 positions of the cysteine pocket (average of 18 and 14.5 different CYPs, respectively). In contrast, for positions outside these regions the average number of different CYPs mentioned per position was 5 (Figure 21 on the following page and supplementary material Table S3 on pages 151-156). 37 (out of 98) positions in the SRSs, 2 (out of 14) positions in the cysteine pocket, but only 4 (out of 407) positions elsewhere were mentioned for more than 18 different CYPs. These 4 positions are functionally relevant, the highly conserved heme interacting residues (standard positions 96 and 100) and two positions at the entrance to the binding pocket (47 and 177).

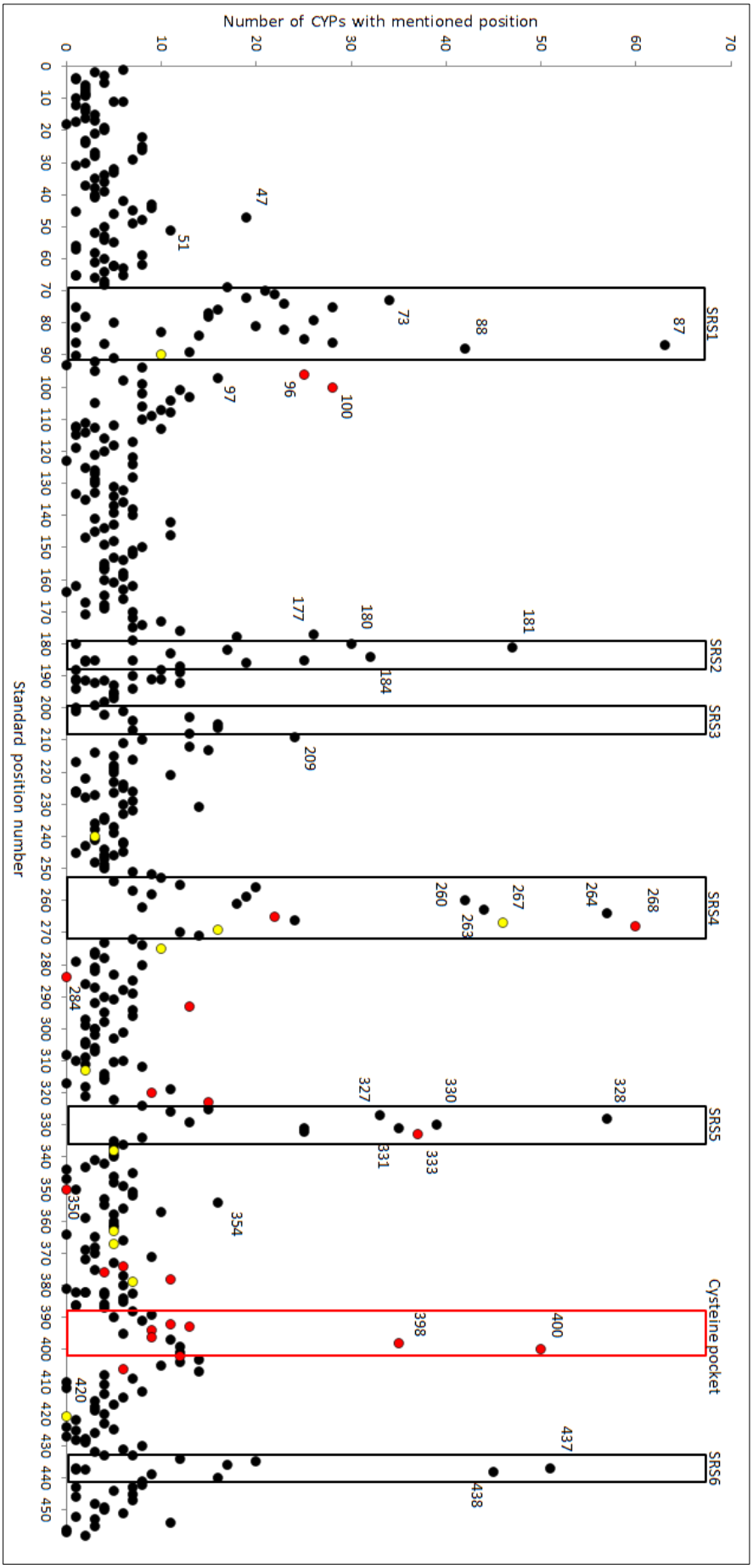


Figure 21: Plot showing the number of CYPs with mentioned standard positions. The substrate recognition sites and the cysteine pocket are marked by boxes. Conserved positions are colored (single conserved amino acids: red, positions where properties such as aromatic, charged, or hydrogen binding are conserved: yellow). Most frequently described positions in each SRS and positions mentioned in the text are labeled by the standard position number.

Positions involved in single nucleotide polymorphisms

Many single nucleotide polymorphisms (SNPs) of CYPs have been associated with diseases, and the rapidly increasing amount of sequencing data reveal new SNPs present in the human population. Therefore, it was of interest to analyze if there are any trends in positions that were mentioned to be related to SNPs. 730 articles mentioned 'polymorphism' or 'SNP' in titles or abstracts together with 611 residues from 47 CYPs (Figure 22 on page 105), corresponding to 301 out of 458 standard positions. In addition, 22 positions in the transmembrane region, which is not covered by the standard numbering schemes, were also described as SNPs. Out of the 18 previously identified conserved residues,[31] 7 were not described to be involved in polymorphisms.

The average number of CYP mentions per standard position was 1.2, within the cysteine pocket it was 1.5, whereas for the SRSs it was 1.4. Thus, from the polymorphism related distribution of the mentioned positions it is not possible to distinguish SRSs, and there are no significant trends suggesting that functionally important regions are more frequently involved in CYP polymorphisms. This might be due to the fact that some of the described polymorphisms do not cause negative effects on CYP function, but are phenotypically neutral. However, by comparing SNP positions and already described functionally relevant positions (like the conserved positions important for CYP function or frequently mentioned SRS positions) using the standard numbering schemes, it should be possible to indicate what biochemical properties of the enzyme will be most probably influenced.

E216 and D301 were described as key determinants of substrate specificity in human CYP2D6, which is involved in metabolism of more than 20% of prescribed drugs.[203] Those residues correspond to the frequently mentioned SRS2 and SRS3 standard positions 184 and 260. Mutations at those positions influence regioselectivity of CYP2D6 toward nitrogen-containing substrates bufuralol and dextromethorphan.[204] R356W and H365Y polymorphisms of CYP21A1 are associated with congenital adrenal hyperplasia and were described to influence CYP activity and stability, but not selectivity.[205] However, the mechanism of the observed functional changes was not discussed. Using the numbering scheme, standard position numbers 325 and 333 were identified for R356 and H365, respectively. Both positions are located on the SRS5. Position 325 is not frequently mentioned and is located at N-terminal region of SRS5, and was described for CYP2B1 to be too far from the binding pocket to influence selectivity.[206] Position 333 is frequently mentioned and is a conserved

propionate-interacting residue.[31] Therefore it can be expected that any change at this position will influence catalytic activity and stability by disturbing heme binding. Though combining literature mining and numbering scheme predicts whether a SNP might have an effect on a biochemical property such as activity or selectivity, the effect of the substitution such as increase or decrease of activity or selectivity cannot be predicted.

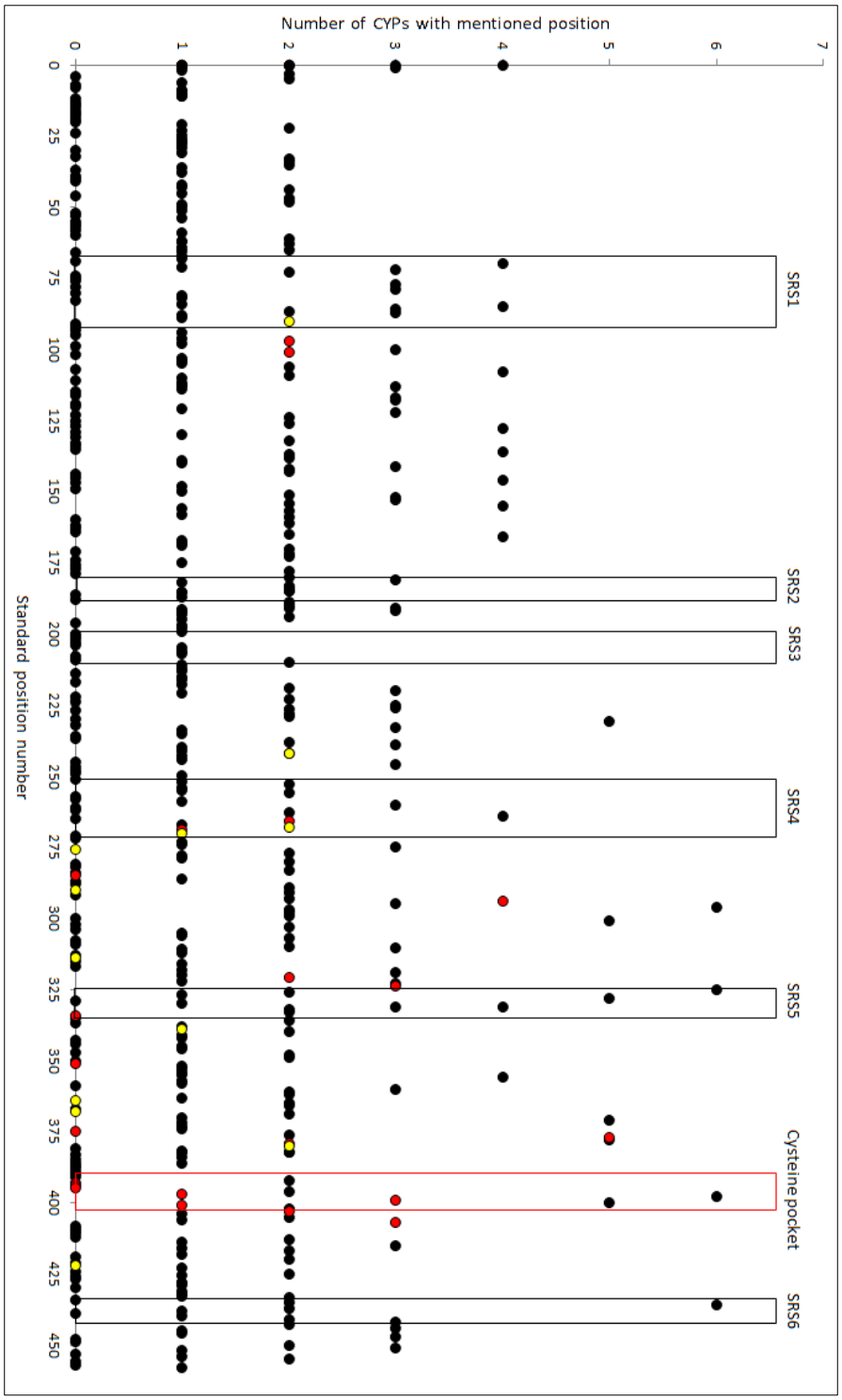


Figure 22: Plot showing the number of CYPs with mentioned standard positions in articles about polymorphisms. The substrate recognition sites and the cysteine pocket are marked by boxes. Conserved positions are colored (single conserved amino acids: red, positions where properties such as aromatic, charged, or hydrogen binding are conserved: yellow).

5.2.3 Discussion

Universal selectivity-determining positions in CYPs

The literature search algorithm allowed for identification of the most frequently mentioned positions in all CYPs (Figure 21 on page 102). Structural correspondence between those positions was possible to identify thanks to application of the standard numbering schemes, which are part of sequence based literature mining algorithm (SBLMA) (Figure 20 on page 100). The majority of these positions were located on the substrate recognition sites (SRSs) and was described to interact with the CYP ligands and therefore to influence selectivity, activity, specificity, or inhibitor binding in a diverse set of CYPs with low global sequence identity. In each SRS we can point to a set of structurally corresponding positions, which were, most frequently mentioned and therefore are good candidates to be described as universal selectivity-determining positions. A prime example is standard position 87 from SRS1, which was mentioned in articles about 63 CYPs. Standard position 87 is the most frequently mentioned position in all CYPs. This position is especially known from CYP102A1 where substitutions of F87 caused significant changes in selectivity towards propylbenzene and terpene substrates, among others.[58,207] Replacement of S122 in CYP1A1 (also standard position 87) by threonine was described to improve 7-methoxy- and 7-ethoxyresorufin O-dealkylase activity.[208] CYP102A1 and CYP1A1 share only 20% sequence identity, which highlights the fact that this position is important in diverse CYPs (Figure 23A on the following page).

The SRS2 is located on the F-G loop and is part of the entrance to the substrate binding pocket. The most frequently mentioned position in this region is standard position 181, which was mentioned in articles about 47 different CYPs. In CYP101A1 (P450cam), mutation T185F at standard position 181 was described to influence its selectivity and to increase the catalytic activity towards norcamphor.[209] Mutation at the corresponding position F206L in CYP2B1 was a major mutation involved in conversion of CYP2B1 from testosterone 16β-hydroxylase to a 15α-hydroxylase.[210] The global sequence identity between CYP101A1 and CYP2B1 is only 17%.

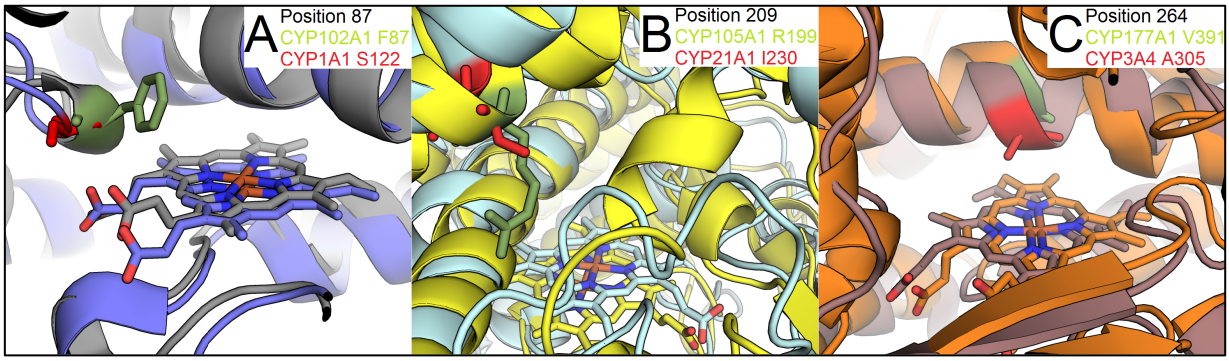


Figure 23: Examples of the structurally corresponding positions universally influencing selectivity among wide range of CYPs. The structures were aligned using PyMOL.[211] A: standard position 87 in gray CYP102A1 residue F87 (PDB code: 1BVY)[43] and in blue CYP1A1 residue S122 (PDB code 4I8V).[212] B: standard position 209 in yellow CYP105A1 residue R199 (PDB code: 2ZBY)[213] and in pale blue CYP21A1 residue I230 (PDB code: 3QZ1).[214] C: standard position 264 in orange CYP177A1 residue V391 (PDB code: 2WIV)[215] and in violet CYP3A4 residue A305 (PDB code: 1W0G).[216]

The SRS3 is an exception among SRSs because there are no frequently mentioned positions in this region. Standard positions 205 and 206 were described in 16 different CYPs (Figure 21 on the page 102) which is significantly less than for most frequently mentioned positions in other SRSs. Position 209 lies on the SRS3 borders described for CYP102A1[80] and was mentioned in articles about 24 different CYPs. In CYP105A1, position R193 (standard position 209) was described to have an important role in vitamin D3 hydroxylation.[213] A mutation at the corresponding position I230 in CYP21A2 was associated with steroid 21-hydroxylase deficiency.[217] CYP105A1 and CYP21A2 only share 22% sequence identity (Figure 23B). A probable explanation to why SRS3 was underrepresented in the frequently mentioned positions is that it lies on the α-helix G, which is not contributing to the binding pocket architecture in the closed form of many CYPs, but it is part of the substrate entrance in the more open CYP structures. Variants at standard position 209 in CYP105A1 and CYP21A1 were described to change activity towards steroids which due to their size occupy not only the binding pocket but also may interact with the α-helix G at the entrance to the binding pocket access channel. Thus, position 209 could be involved in their stabilization and should be included in the SRS3. We speculate that SRS3 might influence the entrance of the substrate to the binding pocket or be relevant only for larger substrates.

SRS4 has the highest number of conserved positions; 4 out of 18 SRS4 positions are conserved (Supplementary materials Table S3 on pages 151-156).[31] SRS4 is located on α-helix I which is the longest α-helix in the CYP-fold and accommodates catalytically relevant residues. The conserved glycine residue at standard position 265 was proposed to be a

structurally important residue and was mentioned in articles about 22 CYPs.[31] Negatively charged residues at standard position 267, threonine at standard position 268 and serine or threonine at the standard position 269, are involved in proton delivery, and were mentioned in articles about 46, 60 and 16 CYPs, respectively.[34,174,175,181] The most frequently mentioned non-conserved position in this region is standard position 264 which was mentioned in articles about 57 different CYPs. V391 (standard position 264) of CYP177A1 (P450 XplA) was described to interact with its inhibitor imidazole.[218] For the mutation A305V at the corresponding position in CYP3A4 a dramatic change in the product profile in progesterone hydroxylation was reported,[219] albeit CYP177A1 and CYP3A4 only share 17% global sequence identity (Figure 23C).

The most frequently mentioned position in SRS5 is standard position 328 which was mentioned in articles about 57 CYPs. This position is located five positions after the EXXR motif and was previously identified by systematic sequence analyses as a selectivity-determining position in most of the CYPs.[59] The corresponding residue L371 in CYP93C2 was described to control accommodation of flavanone in the binding pocket.[220] The corresponding residue L357 in CYP153A6 was described to be determinant of enzymes ω-regioselectivity towards fatty acids.[30] CYP93C2 and CYP153A6 share 18% sequence identity. The most frequently mentioned position in SRS6 is standard position 437, which was mentioned in articles about 51 CYPs. In CYP158A1, the corresponding residue M396 was proposed to serve as a gating residue to the binding pocket.[221] Mutation F494V at the corresponding position in CYP94A2 resulted in a shift of selectivity from ω toward ω-1 hydroxylation of lauric acid.[222] CYP 158A1 and CYP94A2 share 12% sequence identity.

Besides the SRSs, the other most frequently mentioned non-conserved positions are standard positions 177 and 47. Standard position 177, mentioned in articles about 26 CYPs, is located on the α-helix F, 3 positions before the SRS2. However, it is considered to be part of the SRS2 in some human CYPs.[223] Mutation V207T at the corresponding position in CYP3A9 resulted in an improvement of the 6β-testosterone hydroxylase activity.[223] Mutations of the corresponding residue F201 in CYP2C8 resulted in substrate dependent changes in selectivity and activity.[224] Both sequences share only 25% sequence identity. Standard position 47 is located on β-strand B1_2, which is part of a substrate access channel in many CYPs and was described in articles about 19 CYPs. In CYP2C9 and CYP2C19, mutations at the corresponding positions (K72 and R72 respectively) were described to have a significant effect on the enzymatic activity and binding affinity towards tricyclic antidepressant drugs.[225]

In CYP46A1, which shares 22% sequence identity with CYP2C9, substitutions of the corresponding residue L82 were described to influence binding affinity of the cyano- and fluoro-containing drug bicalutamide.[226] Even though standard position 47 is not part of SRSs it seems to have an important role in substrate binding and recognition. Similarly, its close neighbor standard position 51 was described for 11 CYPs and is well known as access channel residue influencing substrate binding in CYP102A1.[80] Another example of non-SRS position influencing selectivity and activity of CYPs is standard position 354. Also located in the access channel, it was described in articles about 16 CYPs. The corresponding residue E433 in CYP5A1 (thromboxane-A synthase) was also described to influence the enzyme's activity.[227] In CYP74D3, the corresponding residue is V379, which was described to change activity from divinyl ether synthases into allene oxide synthases.[228] CYP5A1 and CYP74D3 share 16% sequence identity. This demonstrates that even though SRSs constitute most parts of the binding pocket, they are not exclusively involved in substrate recognition, but there are at least a few other positions that contribute to selectivity in most CYPs.

By systematically analyzing literature and comparing the mentioned positions in the whole CYP family, we found that 14 structurally corresponding positions mentioned in articles about more than 32 CYPs were described to influence selectivity and activity. The most frequently mentioned positions in the SRSs have common effects on substrate binding and activity among a wide spectrum of diverse CYPs, as it was already shown in available literature. Other positions like 209 can play a role in catalysis of a specific class of substrates. Those positions are certainly not the only positions influencing selectivity in CYPs. There might be more positions, which universally influence selectivity of CYPs, but the limited list of the most frequently mentioned positions seems to be a good starting point. Addition of the most frequently mentioned selectivity determining positions to combinatorial libraries of CYPs should be beneficial to find the most diverse set of variants. This could be especially beneficial for CYPs without known structure where the sequence analyses are one of the major rational design methods.

Other universal functionally relevant positions in CYPs

The two most frequently mentioned positions in CYPs outside the SRSs are the heme ligand cysteine at standard position 400 and the heme interacting arginine at standard position 398 in the cysteine pocket.[73,181] In addition, histidine or tryptophan at standard position 96 (in class I and class II, respectively) [31], arginine at standard position 100, and arginine or histidine at

standard position 333 were also frequently mentioned in literature and interact with the heme cofactor in all CYPs.[38,76,229]

Previously, we identified 18 positions with conserved amino acids or amino acid properties (aromatic, charged or hydrogen binding) without described function either in CYP101A1 or CYP102A1.[31] The SBLMA demonstrated that 15 of these positions were also mentioned for other CYPs, except for 3 conserved residues (standard positions 284, 350, and 420) that have not yet been mentioned in the literature (Figure 21 on page 102).

Positions that are involved in CYP-redox partner interactions are of high interest in protein engineering, especially for systems where a CYP domain is coupled with a redox partner from a different organism or even of a different class. In those systems, tuning of the interaction can dramatically improve the catalytic activity.[60] The most frequently mentioned position in class I and class II CYPs redox partner interacting proximal site is standard position 97, mentioned in articles about 16 CYPs. R122 in CYP2B4 (class II) was described as an important residues for the CYP-reductase interaction by which it can influence the electron transfer from human P450 reductase.[88] In CYP3A4 (class II), corresponding residue K127 was described to be involved in the interaction with both cytochrome b_5 and the cytochrome P450 reductase.[83,230] In CYP119 (class I), a mutation at corresponding position D77R was described to improve binding and electron transfer between the CYP and putidaredoxin.[231] The two human CYPs, CYP2B4 and CYP3A4, accept electrons from the same cytochrome P450 reductase, hence it is not surprising that standard position 97 influences electron transfer in both of them. However, CYP119 is a prokaryotic class I CYP accepting electrons from a ferredoxin, and a mutation at the same standard position 97 improved the electron transfer from a non-native class I putidaredoxin. Thus, the CYP-reductase and CYP-redoxin interfaces are overlapping, and some of the residues involved in the interaction might be common for both classes.

Accessibility of literature for automated literature mining

With the rapidly growing number of scientific articles it becomes increasingly important to allow for an automated extraction of information.[232] During the period of 1997-2006, the number of publications in PubMed was growing by 5.5% per year.[233] It was estimated that 20% of published work is freely available as open-access articles.[234] Open-access initiatives such as BioMedCentral and the Public Library of Science (PLoS) encourage open-access publication and invite access via an Application Programming Interface (API) to all of their

articles. While also other journals provide the possibility of open-access publication, not all articles are accessible via APIs which is key for literature mining.[234] API access is rarely provided for articles that are available under an institutional site license agreement only.

In our study, access to the full text articles via APIs was possible for only 22% of all full text articles that have been identified by analyzing the PubMed abstracts. Access via an API provides complete and structured information that can be parsed into a database for further analysis. However, for the majority of all articles, access was only possible via a web interface which is not adequate for systematic data mining. Because of the lack of APIs, literature mining projects such as mining of mutations in protein families[195–197] or of protein-protein interactions[235] have been mainly restricted to analyzing abstracts rather than full text articles. The promise of integrating the loosely connected results of decades of research is finding new insights into sequence-function relationships. But this will only be possible by big data strategies.

5.2.4 Acknowledgements

LG was supported by the People Programme (Marie Curie Actions) of the European Union's 7th Framework Programme (FP7/2007-2013) ITN P4FIFTY under REA Grant Agreement 289217, CV by the DFG in the framework of FOR 1296.

5.3 Redox partner interaction sites in cytochrome P450 monooxygenases: in silico analysis and experimental validation

5.3.1 Abstract

The native redox partners of many novel cytochrome P450 monooxygenases (CYPs) are unknown. Therefore, they are combined with non-native redox partners to obtain catalytically active systems. Understanding the CYP-redox partner interactions is the basis of a successful protein engineering strategy. CYPs are divided into two major classes based on the redox partner type. Class I CYPs accept electrons from iron-sulfur ferredoxin and ferredoxin reductase, class II CYPs from diflavin cytochrome P450 reductase. Here, six redox partner interaction sites (RPISs) were identified by systematic literature, sequence and structure analyses. All six RPISs are proposed to contribute to class II CYP-redox partner interaction interface, whereas four and five contribute to the interaction interface in class I prokaryotic and mitochondrial CYPs, respectively. The significance of identified RPISs was tested by designing mutants. A generic strategy was applied to improve the interactions between CYPs and non-native redox partners. The strategy requires minimal screening efforts. In the fusion system of CYP153A6 (class I) with CYP102A1 reductase (class II), six variants were tested, with mutation K166Q improving electron coupling efficiency from 68% to 89%.

5.3.2 Introduction

Cytochrome P450 monooxygenases (CYPs) are one of the largest protein superfamilies. Despite high sequence diversity, the family exhibits high structural similarity. CYPs usually act as C-H bond oxygenases. Reactions performed by those enzymes are important in many aspects of metabolism across all domains of life.[8] Because of their role in xenobiotic metabolism and carcinogenesis CYPs are of great interest in pharmaceutical industry.[153] Additionally, bacterial, fungi and plant CYPs catalyze many synthetically interesting reactions.[10] Therefore, a generic strategy for the improvement of CYPs catalytic activity is of high interest.

The conserved CYP fold consists of thirteen α–helices and five β–sheets (named A-L and 1-5, respectively)[39] and houses the catalytic heme in a buried substrate binding pocket. The binding pocket is formed by six substrate recognition sites (SRSs), which can be identified in all CYPs by sequence or structure comparison.[40,80] The conserved structural fold allowed to establish two class-specific numbering schemes for CYPs.[31] The numbering schemes allow for identification of structurally corresponding positions in CYPs. This can be done by aligning any given CYP sequence to a structure-guided sequence profiles.[31,71]

To perform oxidation, most of the CYPs require redox partner proteins delivering electrons to the heme.[26] Based on the redox partner type, CYPs were classified into ten different classes.[26] For further simplification, a classification that groups most of the CYPs into two general classes was introduced.[31] Class I CYPs accept electrons from ferredoxins including most of prokaryotic CYPs, mitochondrial CYPs and fusion proteins with ferredoxin as one of the components, while class II CYPs accept electrons from diflavin cytochrome P450 reductases (CPRs) including most of eukaryotic CYPs and fusion proteins with diflavin redox partners. Crystal structures of three CYP-redox partner complexes are known, CYP102A1 with FMN domain of its fusion reductase[43], CYP101A1 with putidaredoxin,[45,46] and CYP11A1 with incomplete adrenodoxin.[188] In all published crystal structures of the CYP-redox partner complexes, the interaction interface is located in a similar region on the CYPs proximal surface.[45,46,188,236] Hence, it is expected that redox partner interaction sites (RPISs) are universal to all CYPs. The identification of the RPISs and residues involved in the redox partner interactions constitutes the basis of a generic strategy for improving catalytic activity and electron coupling efficiency of CYPs. Engineering of the interaction is especially important in CYPs that interact with non-native redox partners. The potential of engineering the redox partner interactions for biotransformation has been demonstrated experimentally, and catalytic activity could be increase up to 5 fold.[60,131]

A previously published review from 2003 by Hlavica et al. provided an initial view on the CYP-redox partner interaction, presented here analyses extend these information by providing details about class-specific differences and identification of regions involved in the CYP-redox partner interactions.[69] Previously, a conservation analysis of all CYPs showed class-specific conservation of heme-interacting residues, but no significant conservation of the CYP-redox partner interaction interface was observed.[31] In this study, we analyzed results from a previous literature mining study[77] to find variants that influence the CYP-redox partner interactions. Amino acid frequency of the described interface positions suggests similar

functional relevance of those positions in all class II CYPs, but different in class I**.** The mutations described in literature to influence the CYP-redox partner interactions were compared between the proteins and six RPISs were identified. The RPISs cover most of the CYPs proximal surface which is consistent with the interfaces from the crystal structures of CYP-redox partner complexes. Based on those results a generic strategy for improving CYP-redox partner interactions between non-native redox partners was proposed. The strategy was validated by testing variants of an artificial interclass fusion of CYP153A6 from *Marinobacter aquaeolei* (class I) with reductase domain of CYP102A1 from *Bacillus megaterium* (class II).[99]

5.3.3 Materials

Computational methods

The previously published version of the Cytochrome P450 Engineering Database (CYPED)[31] was used as a basis for the conservation analyses. Class-specific standard numbering scheme positions were extracted from the CYPED and used for comparisons of the residues between the sequences. Comparisons between the sequences of class I and class II CYPs were conducted based on a conversion table established by a STAMP[94] structural alignment between CYP101A1 and CYP102A1, the reference structures of the standard numbering schemes. Conservation analyses of described in the literature residues influencing CYP-redox partner interaction were performed for sets of class I, class II and human class II CYPs. Class I and class II CYPs were annotated in the database as described previously.[31] 47 human class II CYPs were extracted from the CYPED. Positions described to be important for the CYP-redox partner interactions were found in the results from the previously introduces sequence-based literature mining algorithm.[77] All publications mentioning proximal surface positions were checked to verify whether the position was described to influence the CYP-redox partner interactions. The homology model of CYP153A6 was generated (NCBI gi:120553537), using alignment mode of SWISS-MODEL,[237] the alignment with CYP153A from *Sphingopyxis macrogoltabida* sequence (PDB code: 3RWL) was based on the numbering scheme. PyMOL was used for visualization of protein structures.[238]

Chemicals, enzymes, vectors and strains

Solvents and buffer components were obtained from Alfa-Aesar (Ward Hill, US), Carl-Roth (Karlsruhe, DE), Fluka (Buchs, CH), Macherey-Nagel (Düren, DE) and Sigma-Aldrich (St. Louis, US). 12-hydroxydodecanoic acid was purchased from Sigma-Aldrich. *Pfu*UltraDNA polymerase, *endonucleases, T4 DNA ligase* and isopropyl β-*D*-thiogalactopyranoside (IPTG) were obtained from Fermentas (St. Leon-Rot, Germany). NADPH disodium salt was purchased from Codexis (Jülich, Germany). Glucose-6-phosphate dehydrogenase (1000 U) from *Leuconostoc mesenteroides* was obtained from Roche Diagnostics (Mannheim, Germany). Plasmid pET-28a(+) and *E. coli* strain BL21(*DE3*) originated from Novagen (Madison, Wisconsin, USA). *E. coli* strain DH5α was purchased from Invitrogen (Darmstadt, Germany). Primers were purchased from Metabion International AG (Martinsried, DE).

Engineering of RPIS variants

The construct CYP153A6-CPR containing a 3x(GGS) linker has been described previously.[99] Plasmid pET28a(+) harbouring CYP153A6-CPR was mutated using the QuikChange standard protocol with the primers from Table S4 (on page 157): The resulting PCR products were transformed into competent *E. coli* DH5α cells after *Dpn*I-treatment. Isolated plasmids with the desired mutations (sequencing by GATC-Biotech, Konstanz, Germany) were used to transform competent *E. coli* BL21(DE3) cells. Protein expression, purification and determination of CYP concentration were carried out as described in the previous section.

Protein expression and purification

Cells were grown in shake flasks using TB-medium until an OD_{600} of 0.5-0.7 was reached for induction of protein expression (30°C, 180 rpm). After 16–20 h, cells were collected by centrifugation (7000 rpm, 30 min, 4°C) and resuspended in 50 mM Tris-HCl (pH 7.4). For the determination of CYP concentration and electron coupling efficiency, the RPIS variants were extracted and purified. Resuspended cell pellets were disrupted with the sonifier (Branson Sonifier W250) by using 80 Watt three times for one minute (20 % working interval) at 4°C. The resulting crude extracts were centrifuged (19 000 rpm, 45 min, 4°C), and the supernatants with the soluble proteins were recovered. Protein purification was carried out by anion exchange chromatography with the ÄKTAexplorer (His GraviTrap TALON, GE Healthcare, Freiburg, DE) using a Toyopearl DEAE-650M column (Tosoh, Minato, JP), packed to a volume of 30.4 ml. The column was washed (5 ml min^{-1} working flow) using a linear gradient

protocol with 50 mM of Tris-HCl buffer (pH 7.4) containing 0-1 M of NaCl solution. The elution of the CYP153A fusion proteins occurred at 200-250 mM of NaCl. In addition to the characteristic total protein detection at 280 nm, CYPs were identified by their absorbance at 418 nm (Sligar et al., 1979). This procedure was followed by ultrafiltration using Vivaspin filters with the cut off size of 100 kDa (Vivaspin 100 kDa;Sartorius, Göttingen, Germany). Purified protein solutions were stored in aliquots at -20°C.

Determination of P450 content

*C*oncentrations of the P450 enzymes were determined by the CO differential spectral assay based on the formation of the characteristic Fe(II)-CO complex at 448 nm. Enzymes in cell-free extracts were reduced by the addition of a spatula tip of sodium dithionite, and the CO complex was formed by slow bubbling with CO gas for approximately 30 s. After these steps the sample was incubated for 15 minutes to generate an optimal concentration of the Fe(II)-CO, which has an optimum in absorption by a wavelength of 450 nm. The concentrations were calculated using the absorbance difference at 450 nm and 490 nm (Ultrospec 3100pro spectrophotometer; GE Healthcare), and an extinction coefficient of 91 $M^{-1}cm^{-1}$.[5]

Reaction setup in vitro

The activities and electron coupling efficiency of each CYP153A6-CPR enzyme variant were assayed *in vitro* using dodecanoic acid. Biotransformations were performed using a final volume of 200 µl in 50 mM potassium phosphate buffer (pH 7.5) containing 0.35 µM CYP153A fusion construct, 200 µM NADPH and 200 µM dodecanoic acid (from a 10 mM stock solution in DMSO). The activity analysis was performed by using glucose-6-phosphate/glucose-6-phosphate dehydrogenase (G6P/G6PDH) as cofactor regeneration system. The reaction was started by the addition of enzyme solution. To determine the initial activity, samples were incubated at 30°C and 500 rpm for 5, 10, 15, 30 and 60 min. For the coupling efficiency determination NADPH-consumption of in vitro reactions without cofactor regeneration system was detected until no changes in signal were observed.

Calculation of NADPH-consumption by P450

The slope of the NADPH depletion of the variants was measured in presence and absence of the substrate dodecanoic acid. The difference between the NADPH concentration without substrate and with substrate corresponds to the NADPH consumption of the CYP. Reactions were stopped after no changes in signal were detected and extracted for the analysis of

product formation via GC-FID. The electron coupling efficiencies were determined via the comparison of the NADPH depletion in presence and absence of the substrate.

Sample treatment

Conversion was stopped by adding 30 µl 37% HCl, followed by the addition of internal standard in a final concentration of 0.1 mM decanoic acid. The reaction mixtures were extracted with 0.5 ml methyl tert-butyl ether. The organic phase was collected and evaporated. Samples were resuspended in 45 µl of 1% trimethylchlorosilane in *N*,*O*-bis(trimethylsilyl) trifluoroacetamide and incubated at 70 °C for 30 min for derivatisation.

Analysis of substrates and formed products

Samples were analysed on a GC/MS/FID 7890A instrument (Agilent Technologies, United Kingdom) equipped with a ZB-5 column (30 m×0.25 mm×0.25 µm, Aglient) and with helium as carrier gas (flow rate, 0.69 ml/min; linear velocity 30 cm/s). Mass spectra were collected using electron impact. The injector and detector temperatures were set at 250°C and 330°C, respectively.

For analysis of the substrate and the products, the column oven was set at 140°C for 2 min, raised to 250°C at a rate of 10°C/min, held isotherm for 1 min, and then raised to 320°C at 65°C/min. Reaction products were identified by their characteristic mass fragmentation patterns.[239] Substrate conversions were quantified using calibration curves estimated from a series of standard solutions (0.01 – 1.0 mM fatty acids) treated in the same manner as the samples. Product distributions were calculated from the relative peak areas.

5.3.4 Results

Positions involved in the CYP-redox partner interactions

In each mammalian organism, multiple class II CYPs with high sequence diversity accept electrons from a single, highly homologous cytochrome P450 reductase. This group of CYPs was extensively studied due to their similarity to human CYPs and their importance in drug metabolism. Therefore, previously published literature mining results on CYPs[77] were analyzed to identify proximal surface positions which were described to influence the CYP-redox partner interactions in mammalian CYPs (Table 19 on page 119). Structurally

corresponding positions in the different CYPs were identified using the class II-specific standard numbering scheme,[31] resulting in 27 class II standard positions corresponding to 45 residues found in 10 mammalian CYPs described to influence the CYP-CPR interactions.

To find out if those positions are conserved in a set of proteins accepting electrons from a single redox partner, the sequences of 47 human class II CYPs were analyzed (Table 19 on the following page). Except for the heme-interacting arginine/histidine at standard position 100 and a hydrophobic residue at standard position 112 which were respectively conserved in 92% and 90%, of all class II human CYPs, the frequency of amino acids with specific biochemical properties at these structurally corresponding positions was lower than 50%. This low level of conservation of the interaction site with a single redox partner was unexpected.

The conservation of the 27 redox partner interacting residues was also analyzed in all 3000 class I and 11000 class II CYPs of the CYPED[31] and compared to the conservation inside the subset of 47 human class II CYPs. Except for two outliers (positions 386 and 386.1), the frequency of amino acids with specific biochemical properties at those structurally corresponding positions was very similar in human class II CYPs and differed by only 10% on average. This is despite the fact that class II CYPs interact with diverse reductases, which suggests similar functional significance of the described positions in all class II CYPs. In contrast, conservation of the biochemical properties at the 27 redox partner interacting residues in class I CYPs differed by 21% from human CYPs. This result was expected because class I and II CYPs accept electrons from redox partners with different folds and cofactors, small iron-sulfur cluster ferredoxins and large two domain CPR proteins, respectively.

Redox partner interaction sites in CYPs

The residues described in the literature to be involved in the mammalian CYP-redox partner interaction are localized on six sites of the CYPs proximal surface (Table 19 on the following page and Figure 24 on page 120). Here, we call those sites “redox partner interaction sites” (RPISs). To allow for identification of structurally corresponding positions based on the CYP sequences, the class II standard numbering scheme was used[31] with position numbering derived from CYP102A1 (Table 19). The class-specific standard numbering schemes can be applied to any CYP, and standard positions can be converted between class I and class II using the CYPED web interface (*www.CYPED.BioCatNet.de*).

Table 19: Mammalian CYP class II positions described to be involved in the CYP-CPR interactions. Table consists of class II standard numbering positions (representing structurally corresponding residues), corresponding CYP names and amino acid positions, amino acid property at the standard position, frequency of the amino acid property among human class II, class II and class I CYPs and redox partner interaction site number. Amino acid properties were assigned as follows, positive charge: H, K, R, negative charge: D, E, polar: C, N, Q, S, T, hydrophobic: A, G, I, M, L, P, V, aromatic: F, W, Y, and gaps are indicated by "-".

Class II standard position number	CYP name and amino acid position	Amino acid property	Frequency of the amino acid property in human CYPs	Frequency of the amino acid property in all class II CYPs	Frequency of the amino acid property in class I CYPs	Redox partner interaction site
59	CYP1A1(K94)[82], CYP3A4(K91)[83]	positive charge	82%	60%	43%	1
63	CYP1A1(K99)[82],CYP17A1(K89)[84], CYP19A1(K108)[85]	positive charge	24%	32%	10%	1
65/65.1	CYP2D6(E96)[86],CYP3A4(Y99)[83], CYP3A4(C98)[87]	negative charge /aromatic/polar	43%/6%/28%	30%/6%/32%	20%/1%/13%	1
68	CYP1A1(K105)[82]	positive charge	16%	9%	7%	1
97	CYP2B4(R122)[88], CYP3A4(K127)[83]	positive charge	59%	67%	33%	2
99	CYP1A1(R135)[82]	positive charge	35%	22%	21%	2
100	CYP1A1(R136)[82], CYP3A4(R130)[83], CYP2B1(R125)[89], CYP1A2(R137)[90]	positive charge	92%	94%	98%	2
101	CYP1A1(R137)[82], CYP2B4(R126)[88], CYP2C9(R125)[91]	positive charge	69%	67%	54%	2
104	CYP3A4(S134)[83]	polar	37%	45%	57%	2
108	CYP2B4(R133)[88], CYP2D6(R140)[86]	positive charge	59%	29%	17%	2
110	CYP2B4(F135)[88], CYP21A1(R132)[92]	aromatic/ positive charge	14%/10%	13%/16%	1%/61%	2
112	CYP2B4(M137)[88]	hydrophobic	90%	74%	91%	2
113	CYP2B4(K139)[88], CYP3A4(K143)[83]	positive charge	57%	50%	30%	2
236	CYP2B4(V267)[93]	hydrophobic	55%	53%	50%	3
239	CYP2B4(V270)[93]	hydrophobic	51%	23%	27%	3
305	CYP3A4(Y347)[83]	aromatic	47%	38%	-	4
310	CYP17A1(R347)[84]	positive charge	43%	26%	-	4
319	CYP17A1(R358)[84]	positive charge	73%	54%	17%	4
383	CYP2B4(L420)[93], CYP19A1(K420)[85]	hydrophobic/ positive charge	8%/16%	30%/13%	-	5
386/386.1	CYP1A1(K440)[82], CYP2B4(R422)[88],	positive charge	55%/63%	17%/5%	-	5
388	CYP3A4(Y430)[83]	aromatic	24%	40%	-	5
397	CYP1A1(K453)[82], CYP2B4(K433)[88], CYP2D6(R440)[86]	positive charge	51%	45%	7%	6
399	CYP1A1(R455)[82]	positive charge	12%	10%	29%	6
404	CYP3A4(R446)[83]	positive charge	31%	26%	27%	6
407	CYP1A1(K463)[82], CYP2B4(R443)[88], CYP19A1(R449)	positive charge	41%	24%	62%	6

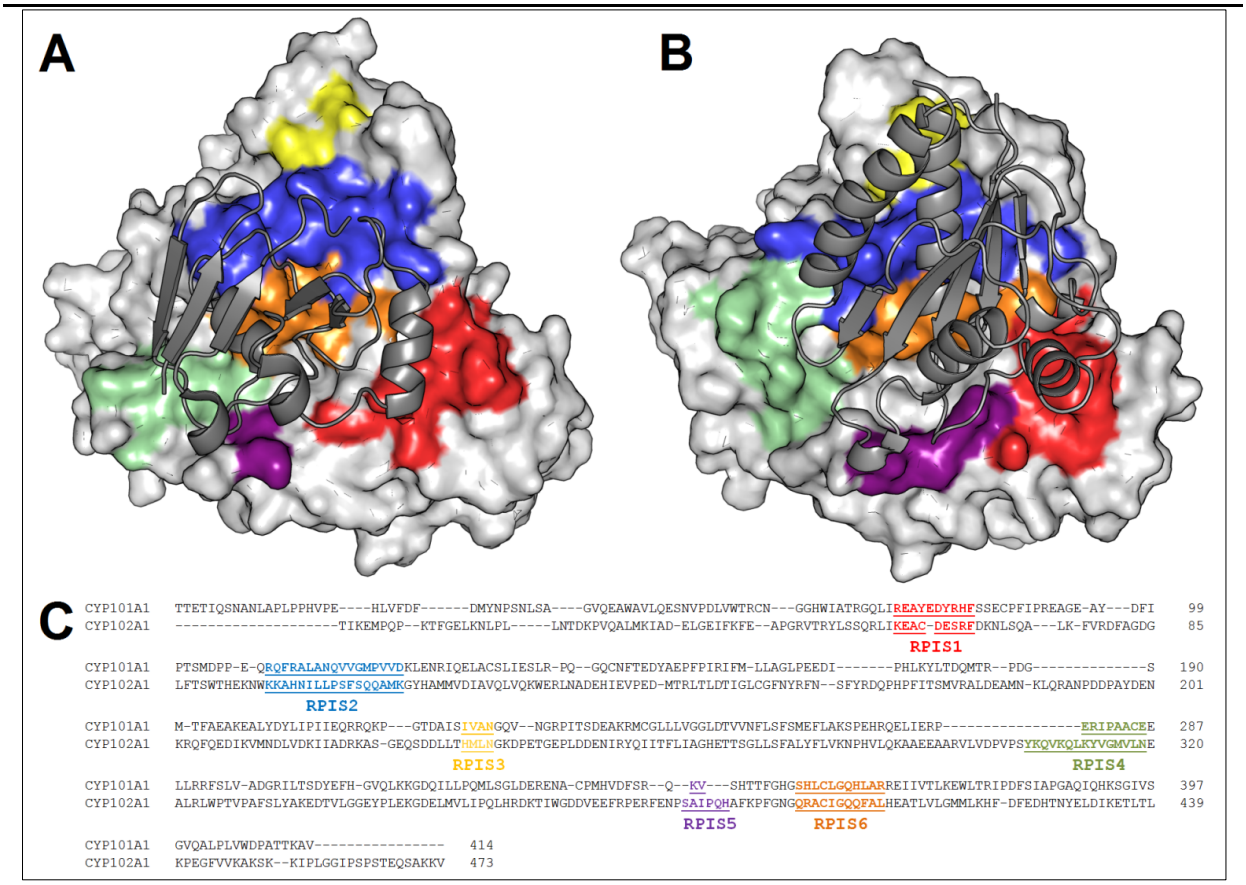


Figure 24: Reductase interaction sites on the proximal surfaces and sequences of CYP101A1 (PDB code: 3W9C[45]) and CYP102A1 (PDB code:1BVY[43]). The RPISs are highlighted in different colors: RPIS1 red, RPIS2 blue, RPIS3 yellow, RPIS4 green, RPIS5 purple, and RPIS6 orange. A: CYP101A1 structure with co-crystalized putidaredoxin (showed as grey cartoon), which rests over most of the RPISs. B: CYP102A1 structure with co-crystalized FMN domain of the reductase (showed as grey cartoon), which rests right over the RPISs. C: sequence alignment of a STAMP[94] structure alignment between CYP102A1 (PDB code: 1ZOA)[95] and CYP101A1 (PDB code: 1PHG)[96] with marked RPISs.

RPIS1 (class II standard positions 59 – 67) is located on the α-loop-β element formed by α-helix B and β-strand 1_5. Positively charged, negatively charged, polar and aromatic residues from this region were described to be involved in the interaction.

RPIS2 (class II standard positions 97 – 113) is located on the α-loop-α element formed by α-helix C and the N-cap of α-helix C'. Positively charged, hydrophobic, polar and aromatic residues from this region were described to be involved in the interaction. RPIS2 contains a conserved arginine/histidine at position 100, which was described to be involved in electron transfer and heme stabilization.[229]

RPIS3 (class II standard positions 236-239) is located at the C-cap of α-helix H. This is the shortest RPIS and only hydrophobic residues from this site were described to contribute to the CYP-CPR interactions.

RPIS4 (class II standard positions 305-319) is located on the α-loop-α element formed by the N-cap of α-helix H and α-helix J'. Positively charged and aromatic residues from this region were described to be involved in the interaction.

RPIS5 (class II standard positions 383-388) is located on the cysteine pocket and part of the preceding loop. Positively charged, hydrophobic and aromatic residues from this region were described to be involved in the CYP-CPR interaction.

RPIS6 (class II standard positions 397-407) is located on the cysteine pocket and the N-cap of α-helix L. Positively charged residues were described to be involved in the CYP-CPR interaction in this region. RPIS6 contains multiple conserved residues, a heme propionate interacting arginine at position 398, the heme ligand cysteine at position 400, a conserved glycine at position 402 and a conserved alanine at position 406.[31]

The RPISs cover most of the proximal surface of CYPs which is the interaction site for class I and class II redox partners. Thus, positions in these regions are expected to influence the CYP-redox partner interactions in most CYPs (Figure 24). However, the difference in size between class I ferredoxins (~105 amino acids) and class II FMN-domains (~150 amino acids) suggests that not all RPISs might be involved in class I CYP-redox partner interactions as compared to class II CYPs. An additional region between RPIS5 and RPIS6, near to the cysteine pocket (positions 389-396) has not been described yet to include important residues for CYP-CPR interaction. This region is also functionally and structurally significant, because four out of seven its positions are conserved.[31] P392, G394 and G396 are structurally relevant positions to maintain the shape of the cysteine pocket, and F393 was described to interact with the heme.[183] None of the remaining three positions of this region were described to influence the CYP-redox partner interactions in mammalian class II CYPs.

Re-designing RPISs of CYP153A6

The identified RPIS positions are expected to influence CYP-redox partner interactions and thus electron coupling efficiency and catalytic activity in most of the CYPs. To test this assumption, we designed single mutants of a previously described interclass fusion construct containing class I CYP153A6 from *Marinobacter aquaeolei* and the reductase domains of

class II CYP102A1 from *Bacillus megaterium*.[99] While most of the previous studies on CYP-redox partner interaction sites have been performed by extensive alanine scanning or site-saturation mutagenesis, our design strategy had two goals: to demonstrate that RPISs are hotspots of CYP-redox partner interactions, and to design a minimal library of mutants with improved electron coupling efficiency. Our strategy is generic and can be applied to any non-native CYP-redox partner pair. This approach is directed at making the proximal surface of any selected CYP more similar to the one of the reductases' natural recipient of electrons, and is done by mimicking charge composition of the RPISs. In total, six positions were selected where the amino acid charge differed in CYP153A and CYP102A1. Charged residues were selected because of the reported high significance in the CYP-redox partner interactions.[82,97,98] The selected CYP153A6 residues are structurally corresponding to the RPIS positions described in the literature for mammalian CYPs (Table 19 on page 119).

Residues at the RPIS positions were compared between CYP153A6 and CYP102A1 (Table 20 on the following page), based on the standard numbering scheme. CYP102A1 is the reference for class II numbering, and its amino acid positions were compared to structurally corresponding CYP153A6 positions based on class I numbering scheme position assignment. Class I positions were converted to class II using previously published numbering scheme conversion table.[31] Six positions introducing a change in charge were selected for mutagenesis: L115K and S120D in RPIS1, D153K and K166Q in RPIS2, R422Q and E425L in RPIS6. Variants S122D and D125S were omitted because they are located further away from the center of the proximal surface than S120. Variant R399R was omitted because it is a neighbor of the proximal cysteine and might disturb CYPs catalytic activity. No mutations were introduced in RPIS3 because it has been described to be important for hydrophobic interactions and the proposed variants in this region did not introduce such residues. RPIS4 and RPIS5 lie on regions which are shorter in class I than in class II CYPs and a reliable comparison between CYP153A6 and CYP102A1 in those regions was not possible.

Table 20: Comparison of amino acid residues at the CYP-CPR interaction interface between CYP102A1 and CYP153A6. Table consists of class II standard position numbers, amino acids at the corresponding positions in CYP102A1 and CYP153A6, position numbers in CYP153A6 and redox partner interaction site number. Positions selected for mutagenesis are in bold and highlighted grey, CYP153A6 amino acids were replaced by CYP102A1 amino acids.

Class II standard position number	CYP102A1	CYP153A6	CYP153A6 position number	Redox partner interaction site
59	**K**	**L**	**115**	**1**
63	**D**	**S**	**120**	**1**
65	S	D	122	1
68	D	S	125	1
97	**K**	**D**	**153**	**2**
99	A	Q	155	2
100	H	R	156	2
101	N	S	157	2
104	L	Q	160	2
108	S	A	164	2
110	**Q**	**K**	**166**	**2**
112	M	L	168	2
113	K	K	169	2
236	H	L	280	3
239	N	S	283	3
305	Y	-	GAP	4
310	Q	-	GAP	4
319	L	S	345	4
383	S	-	GAP	5
386	P	-	GAP	5
388	H	-	GAP	5
397	Q	V	415	6
399	A	R	417	6
404	Q	R	422	6
407	**L**	**E**	**425**	**6**

Experimental characterization of CYP153A6 variants with changed RPISs

Catalytic activity and electron coupling efficiency of the designed variants and wild type (WT) CYP153A6 chimera were measured to quantify the influence of the RPIS mutations on the conversion of dodecanoic acid. It was not possible to express variant R422Q. Therefore, the five expressed variants and the wild type were subjected to a FPLC purification system

with anion exchange column (AEC), and tested *in vitro* for electron coupling efficiency and catalytic activity (see supporting information).

The determined electron coupling efficiency of the WT (67.8%) was in good agreement with the literature.[99] Three RPIS variants showed an increase in electron coupling efficiency: S120D (72.6%), D153K (76.2%) and K166Q (89.3%) (Table 21). The analysis of initial reaction rates and conversion after 1h showed that the WT was more active than the RPIS variants. The best variants K166Q and S120D showed 27% and 37% lower conversion then the WT after one hour, respectively. The other three variants showed more than 50% lower than the WT conversion after one hour.

Table 21: Initial reaction rates, conversion after 1h and coupling efficiency of wild type CYP153A6-CPR and RPIS variants. The highest values for each property are in bold.

	WT	L115K	S120D	D153K	K166Q	E425L
Initial rate [µmol/min*µmol]	**23.2±0.7**	1.9±0.1	5.3±0.7	2.2±0.2	10.6±0.6	2.2±0.6
Conversion after 1 h [%]	**80.1***	27.4±1.7	44.4±6.0	27.2±0.4	52.8±3.7	29.8±6.1
Coupling efficiency [%]	67.8±11.5	56±1.4	72.6±4.5	76.2±5.4	**89.3±6.9**	63.1±2.8

*one sample analyzed

5.3.5 Discussion

Redox partner interaction sites in class II CYP systems

Improving the interactions between cytochrome P450 monooxygenases and their redox partners is a major challenge in CYP engineering. Modifications of the CYP-redox partner interaction interface were demonstrated to improve electron coupling efficiency and catalytic activity.[60,131,132,240] Thus, the relevance of redox partner interaction sites (RPISs) for CYP activity and electron coupling efficiency is comparable to the relevance of the substrate recognition sites (SRSs) for CYPs selectivity and specificity.[40] A comprehensive analysis of published data on the mammalian CYP-CPR interaction, allowed for identifying six major RPISs on the proximal surface of the CYPs. Positions at those sites have been shown to influence CYP-redox partner interactions in different CYPs. Interestingly, those positions are not conserved in class I and class II CYPs[31] and not even in human class II CYPs which accept electron from a single reductase.

Interactions of class II CYPs with their native redox partner were mainly studied in mammalian CYPs (Table 19 on page 119), but the RPISs identified in mammalian CYPs are

also part of the interaction interface in other class II CYPs and fusion proteins including the class II redox partner homologues. The first crystal structure of a CYP-CPR complex, CYP102A1 in contact with its FMN domain, provided many insights into the CYP-redox partner interface.[43] CYP102A1 is a natural fusion protein containing a heme domain and a cytochrome P450 reductase domain (CPR). CYP102A1 is also the reference protein for the class II numbering scheme, therefore class II standard numbering positions are corresponding to this CYPs position numbering. Another class II system for which the CYP-redox partner interactions were studied in detail is CYP6AB3 from parsnip moth and its non-native redox partner, a CPR from house fly. Positions that were described to influence the interaction or change the catalytic activity of those enzymes have been described for all RPISs. For CYP6AB3 and CYP102A1, RPIS1 standard positions 62 and 64, respectively, were mutated in variants with improved catalytic activity (V92A[131] and E64G[241], respectively). For CYP102A1, RPIS2 standard positions H100, N101 and L104 were described as crucial for CYP-CPR interaction[242] and these positions are structurally corresponding to positions described for mammalian class II CYPs (Table 19). RPIS3 standard position H236Q was included in variants resulting from directed evolution of CYP102A1 used for conversion of ethane to ethanol (mutation H236Q),[128] this position also influenced CYP-redox partner interaction in CYP2B4.[93] Several RPIS4 positions were part of combinatorial libraries of CYP102A1[80], and mutation at standard position Q307H improved its catalytic activity towards diclofenac, ibuprofen, and tolbutamide.[243] The complete RPIS5 of CYP102A1 was described to interact with the FMN domain [242], and variants including mutation at standard position P386S significantly increased the catalytic activity of CYP102A1 towards β-ionone.[244] The structurally corresponding residue also influenced CYP-redox partner interactions in mammalian CYPs (Table 19). In RPIS6 of CYP102A1, mutation at standard position I401P was described to change the redox potential of heme and resulted in an increase of the first electron transfer rate by 10%,[245] demonstrating that engineering the CYP-redox partner interface can improve the electron transfer and thus the physicochemical properties of the enzyme. Changes of the redox potential are frequently observed when targeting RPIS5, RPIS6, or the heme-interacting cysteine pocket which is positioned between RPIS5 and RPIS6.[245,246]

The literature data for class II CYPs showed that mutations in the six RPISs can improve catalytic activity and electron coupling efficiency. Most interestingly, many of the positions are also involved in interactions between class I CYPs and their redox partners, and as shown

here also in the interclass systems. Positions of the RPIS1, RPIS2, RPIS5, and RPIS6 were most frequently described to influence the interaction. This underlies their importance in CYP-redox partner interaction, and might be connected to their position in the center of the proximal surface and in proximity to the heme. This is similar to the substrate recognition sites, where SRS1 and SRS5 have a major role in determining selectivity due to their position close to the heme.[58] And similar to R47 or Y51 in CYP102A1, which influence selectivity despite not being part of a SRS, there are positions outside of the RPISs that influence the CYP-redox partner interactions, such as two residues located between RPIS5 and RPIS6 in CYP101A1 (H352 and G353).[80] It is interesting to note that many mutations in the RPISs resulting in improved catalytic activity involve uncharged residues, which demonstrates that CYP-redox partner interactions is more than pure electrostatic interactions.

Redox partner interaction sites in class I CYP systems

Class I CYPs accept electrons from ferredoxins, which are considerably smaller than CPRs, the redox partners of class II CYPs. Therefore, systematic differences between the CYP-ferredoxin and the CYP-CPR interfaces, which also reflected structural differences between prokaryotic class I and class II CYPs are expected.[72] Part of the RPIS4, α-helix J' is not present in most of prokaryotic class I CYPs, similarly length of the RPIS5 is also reduced in those enzymes (Figure 24 on page 120). Apart from these differences, all six RPISs are present in class I and class II CYPs in a similar extent and can be directly compared: RPIS1 (positions 72-82 or 59-67 in the class I or class II numbering scheme, respectively), RPIS2 (positions 109-125 or 97-113), RPIS3 (positions 222-225 or 236-239), RPIS4 (positions 279-286 or 305-319), RPIS5 (positions 344-345 or 383-388), and RPIS6 (354-364 or 397-407) (Figure 24C). CYP101A1 is the most extensively studied class I CYP, and a model enzyme for understanding CYP catalytic mechanism.[181] CYP101A1 is also the reference protein for class I numbering scheme, therefore class I standard numbering positions are corresponding to this CYPs position numbering. Recently, the structural basis of the interactions between CYP101A1 and its redox partner putidaredoxin was revealed.[45,46] Before the X-ray structure of the complex became available, a set of charged residues was predicted to mediate CYP-redox partner interactions and electron transfer: R72 (RPIS1), R112 (RPIS2), K344 (RPIS5), and R364 (RPIS6). These positions correspond to class II standard positions 59, 100, 384, and 407, respectively, which were described to influence CYP-redox partner interactions in class II CYPs (Table 19 on page 119).[98,247] The crystal structures of the CYP101A1-putidaredoxin complex extended the number of residues that are involved in the interaction: E76 (RPIS1),

R109, A113, N116, M121, P122, and R125 (RPIS2), H352 and G353 (between RPIS5 and RPIS6), and L356 L358, Q360, and H361 (RPIS6), corresponding to class II standard positions 62.1, 97, 101, 104, 109, 110, 113, 395, 396, 399, 401, 403, 404, respectively. 50% of these positions were already described to influence class II CYP-redox partner (Table 19 on page 119).[45,46] RPIS3 and 4 were not described to be involved in the interaction between CYP101A1 and putidaredoxin, which is probably caused by the lack of α-helix J' and the smaller size of the putidaredoxin in comparison to the class II redox partner FMN domain (Figure 24 on page 120). Interestingly, residues between RPIS5 and RPIS6 were described to be interacting with the putidaredoxin, which is a consequence of the different architecture of the cysteine pocket in class I and class II CYPs,[31,72] suggesting that RPIS5 and RPIS6 might be merged in class I CYPs.

CYP-ferredoxin interactions were successfully improved in different prokaryotic CYP systems pointing to common interaction hotspots on the RPIS2. For many CYPs, interactions with non-native redox partners were improved by mutating structurally corresponding residues in RPIS2 (standard positions 120 and 108 in class I and class II CYPs, respectively). Mutations of CYP105 (P450moxA) from *Nonomuraea recticatena* (T115A),[132], CYP105A3 from *Streptomyces carbophilus* (T119S),[60] and CYP197A from *Pseudonocardia autotrophica* (T107A)[240] improved the electron transfer in class I CYPs, and corresponding positions were also described to influence CYP-CPR interactions in class II CYPs (Table 19). Another RPIS2 mutation, D77R (class I standard position 109, class II standard position 97) improved binding between CYP119 from *Sulfolobus solfataricus* and its non-native redox partner putidaredoxin.[231] This position also influenced CYP-redox partner interactions in class II CYPs, and is structurally corresponding to the K166Q which was experimentally investigated variant with the highest electron coupling efficiency in this paper for the CYP153A6 fusion protein. Mutation N363Y in RPIS6 (class I standard position 361, class II standard position 404) was part of a variant with improved electron transfer efficiency between putidaredoxin and CYP105A3.[60]

Studies on the interaction between human mitochondrial class I CYPs and their redox partner adrenodoxin point to RPIS1, RPIS2, RPIS4, and RPIS5 as the most significant sites of the interaction. In CYP27B1 mutation G102E in RPIS1 (class I standard position 78, class II standard position 64) resulted in 80% decrease in catalytic activity.[248] In CYP11A1, mutations K148A and K149A in RPIS2 (class I standard positions 109 and 110, class II standard positions 97, 98) resulted in 1.5-fold improved activity in pregnenolone biosynthesis.[249] In

CYP27A1, Mutations of K354A, K358A, and R418S in RPIS4 and RPIS5 (class I standard positions 282, 286 and 344, class II standard positions 315, 319, and 384) decreased adrenodoxin binding.[97] Because human mitochondrial class I CYPs have α-helix J', in contrast to prokaryotic class I CYPs, RPIS4 plays more significant role in those enzymes.

Even though class I and class II CYPs accept electrons from redox partners of a different fold and show some structural differences, most of the RPISs identified for class II CYPs are also found in class I CYPs. In addition, some positions were described to be involved in the CYP-redox partner interactions in both classes of CYPs (class II standard positions 59, 64, 97, 100, 101, 104, 108, 110, 113, 319, 399, 401, 404, 407). Based on the literature data about engineering of CYP-ferredoxin interactions in prokaryotic CYPs, RPIS3 and RPIS4 do not seem to play significant role. Class I RPIS2 standard position 120 is especially significant in the interaction between class I prokaryotic CYPs and ferredoxins. In human class I CYPs, the significance of RPISs is similar, additionally because of presence of α-helix J', RPIS4 is also involved in adrenodoxin binding.

Protein engineering strategies for improvement of CYP-redox partner interactions

To test the relevance of the RPIS we analyzed the literature information to design RPIS variants of the CYP153A6 chimera.[99] CYP153A6 from *Marinobacter aquaeolei* is a class I CYP accepting electrons from a ferredoxin. However, in the chimera it is accepting electrons from the reductase domain of CYP102A1 from *Bacillus megaterium,* a class II CYP. The designed mutations were located on RPIS1, RPIS2, and RPIS6. All five tested variants exhibited changed catalytic activity, and most of them had changed electron coupling efficiency in the oxidation of dodecanoic acid (Table 21 on page 124). The best variant K166Q (class II standard position 110) had an electron coupling efficiency which was 22% higher than the wild type chimera, while its catalytic activity was decreased by 35%. In the variant, a positively charged surface residue was exchanged for a neutral residue which probably improved charge pairing between the CYP and its redox partner. Our results demonstrate that positions frequently mentioned in literature to influence catalytic activity and electron coupling efficiency in different CYPs (Table 19) are promising candidates for the design of highly enriched mutant libraries. Moreover those positions can be transferred also to class I CYPs to engineer fusion proteins with CPR-like reductases, underlining their universal importance. The design and characterization of a small number of CYP153A6 variant was

intended as a feasibility study. We expect that screening of a larger RPIS library will result in further enzyme variants with improved properties.

The proposed protein engineering strategy can be transferred to other class I or class II CYPs by applying the standard numbering scheme for class I and class II CYPs.[31] The strategy can be extended in several ways. We have shown that mimicking the surface charge of the natural CYP partner is promising, but in many examples in literature the electron transfer was improved by mutations not involving charged residues. Therefore, the design of mimicking mutations of neutral residues might be more advantageous. A similar strategy targeting the electron transfer route was previously used to improve catalytic activity of a fusion protein, [60] and the putative residues of the electron transfer path were exchanged to the corresponding residues of the natural electron acceptor. A more extensive strategy would be to exchange all frequently described RPIS positions to the most frequently occurring residues. This approach would benefit from the existence of universal hotspots such as standard position 120 in class I CYPs (corresponding to standard position 108 in class II).

Because electron uncoupling causes the production of hydrogen peroxide and other reactive oxygen species which will lead to irreversible inactivation of the enzyme,[22] improving the coupling efficiency increases stability of the CYP system and has a considerable benefit in whole cell biotransformations.[23] Therefore, the CYP153A6 mutant K166Q designed in this study will be especially useful for whole cell biotransformations.

5.3.6 Acknowledgements

L.G. received funding from the People Programme (Marie Curie Actions) of the European Union's 7th Framework Programme (FP7/2007-2013) ITN P4FIFTY under REA Grant Agreement 289217 and S.M.H. received funding for research, technological development and demonstration under grant agreement n° 613849.

6. Supporting Information

6.1 Modeling of CYP101A1 variants stereoselectivity towards methylated ethylbenzene derivatives

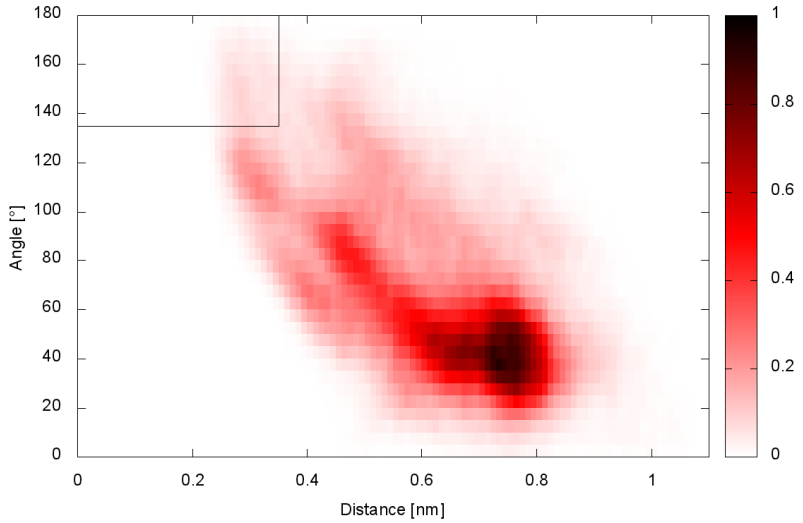


Figure S1: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant Y96F and 1-ethyl-2-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

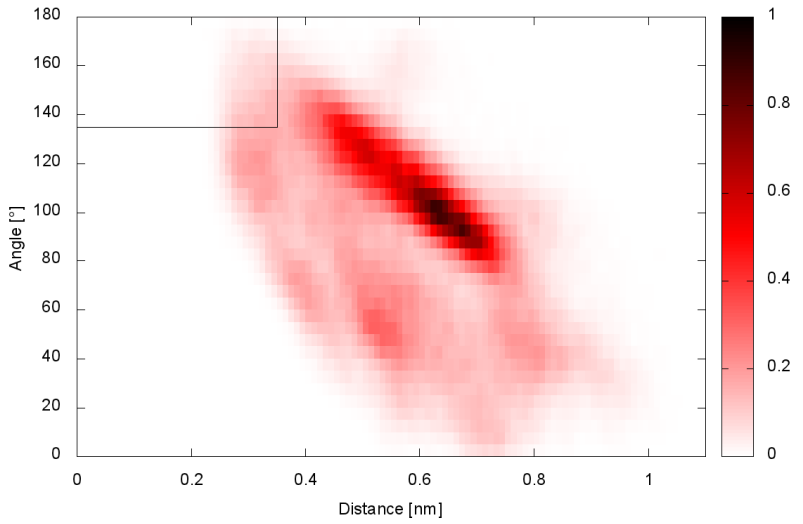


Figure S2: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant Y96F and 1-ethyl-2-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

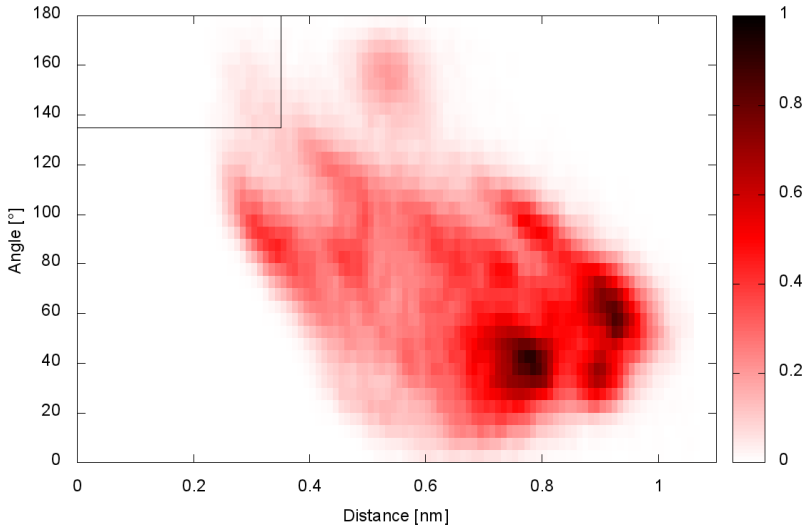


Figure S3: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant Y96F and 1-ethyl-3-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

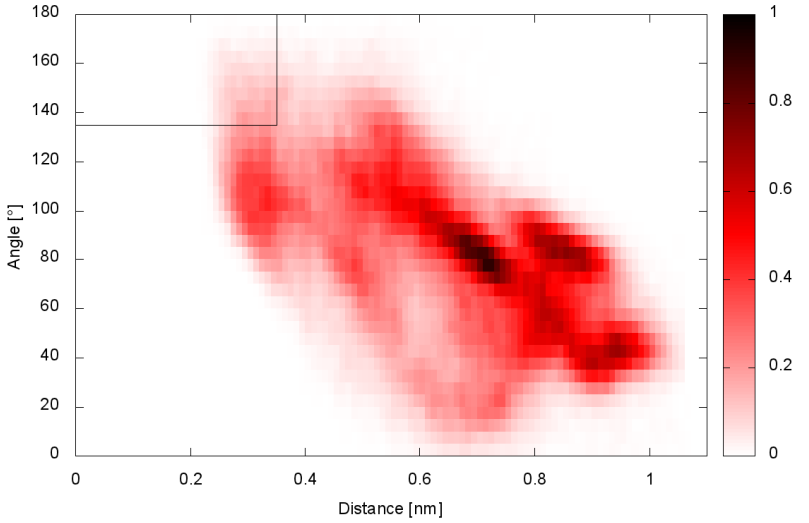


Figure S4: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant Y96F and 1-ethyl-3-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

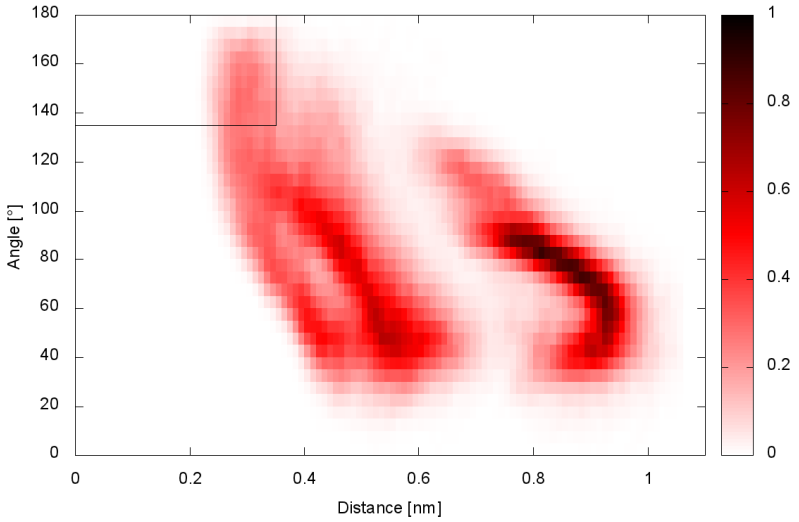


Figure S5: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant Y96F and 1-ethyl-4-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

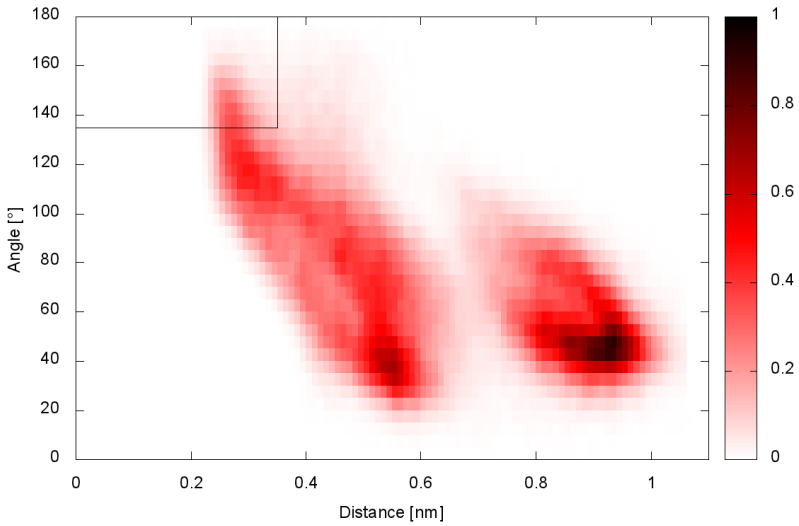


Figure S6: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant Y96F and 1-ethyl-4-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

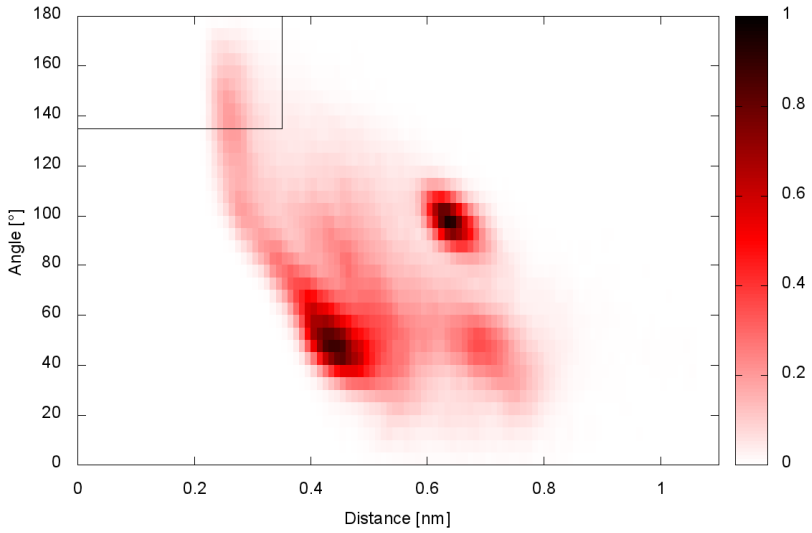


Figure S7: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant VF and 1-ethyl-2-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

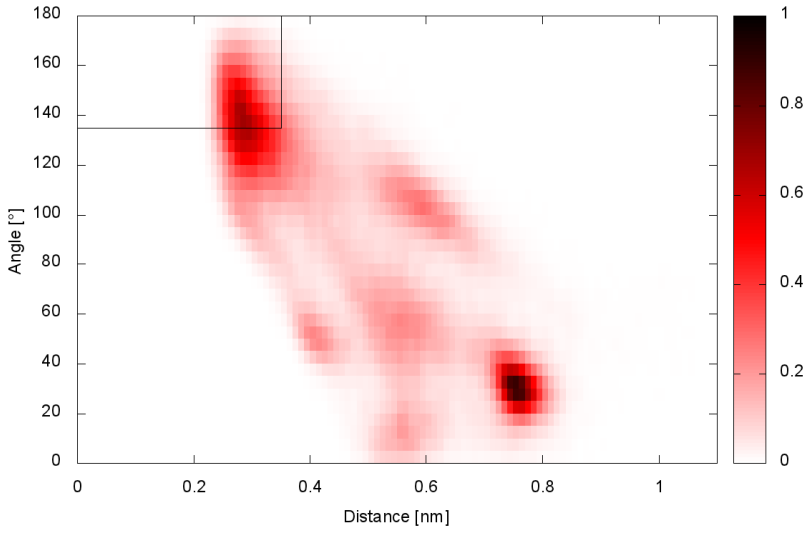


Figure S8: Heat map representing distance dH-O and angle αC-H-O values for pro-s near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant VF and 1-ethyl-2-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

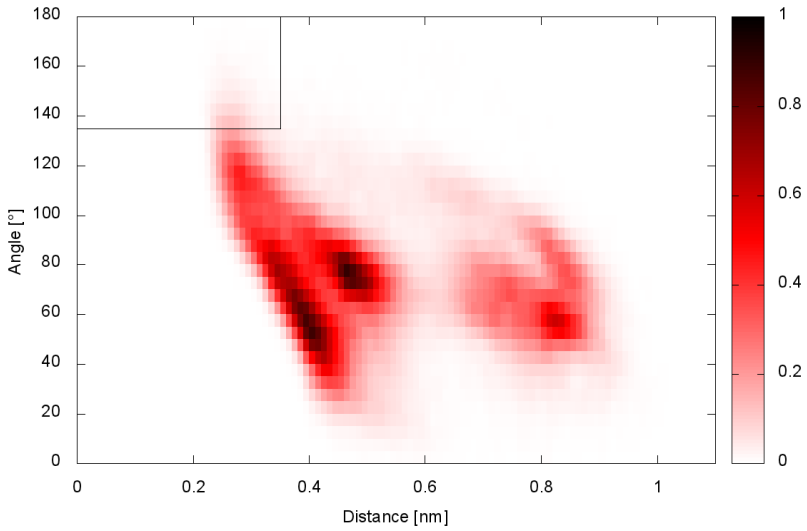


Figure S9: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant VF and 1-ethyl-3-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

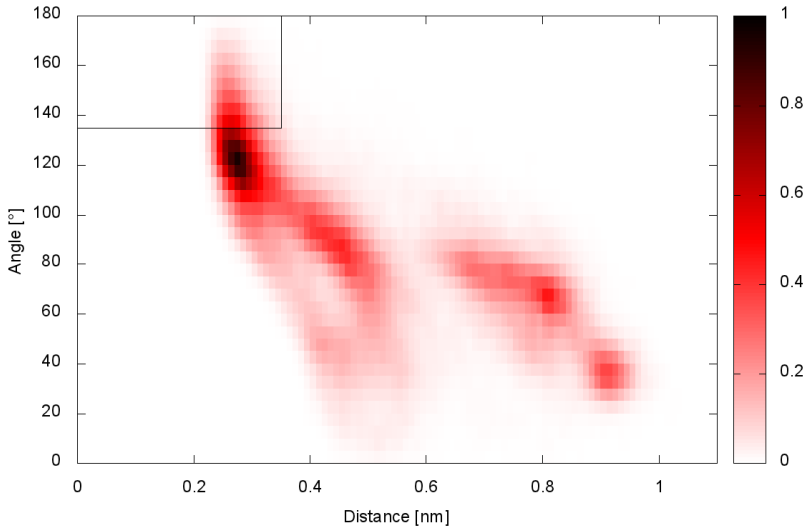


Figure S10: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant VF and 1-ethyl-3-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

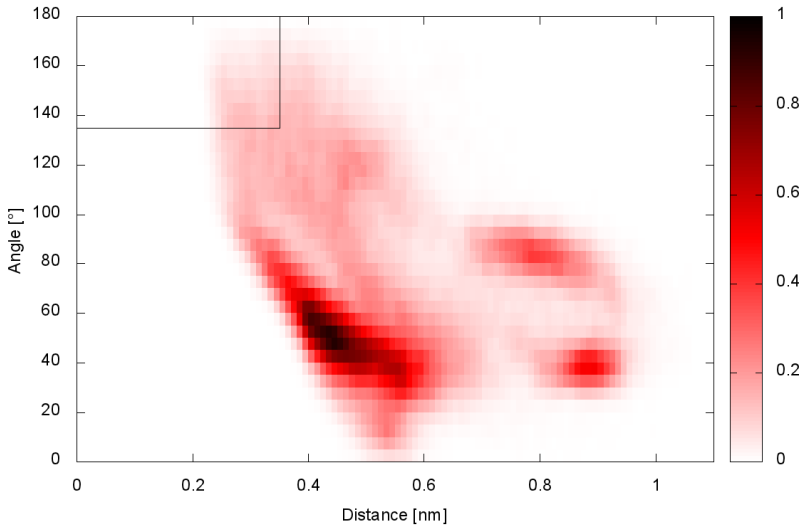


Figure S11: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant VF and 1-ethyl-4-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

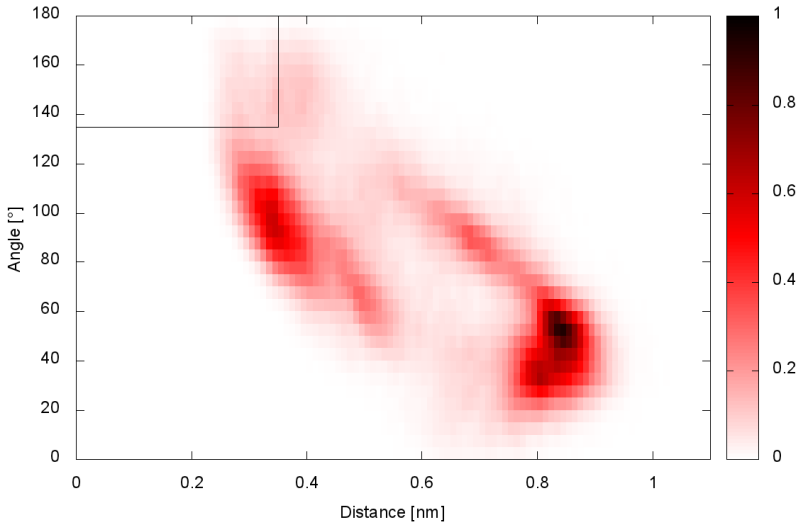


Figure S12: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant FL and 1-ethyl-2-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

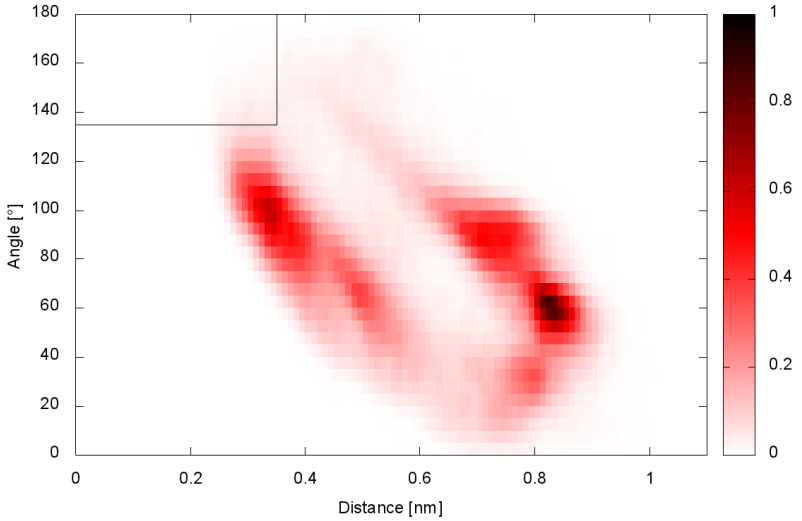


Figure S13: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant FL and 1-ethyl-2-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

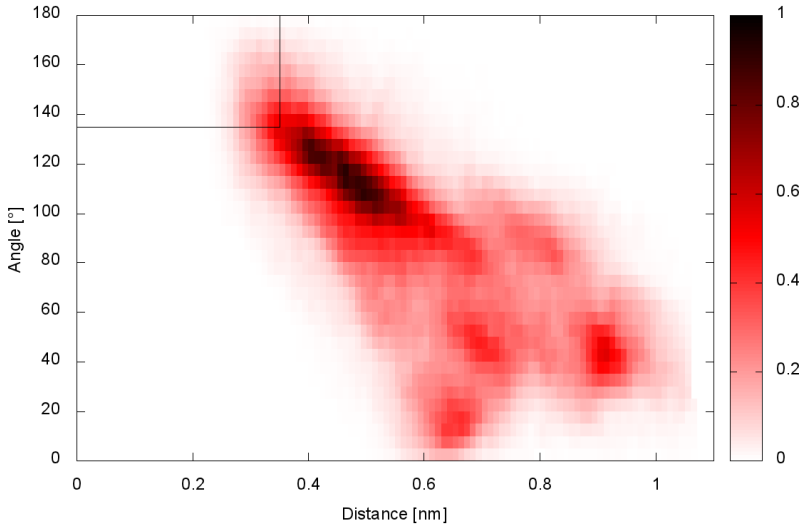


Figure S14: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant FL and 1-ethyl-3-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

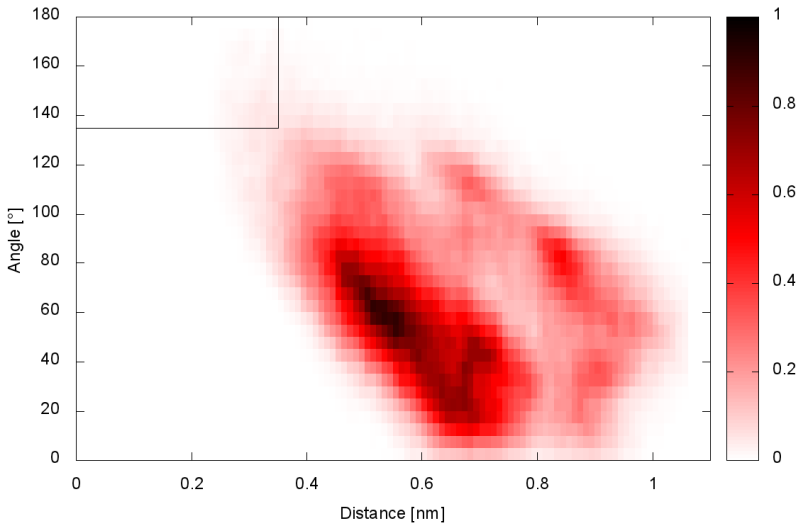


Figure S15: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant FL and 1-ethyl-3-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

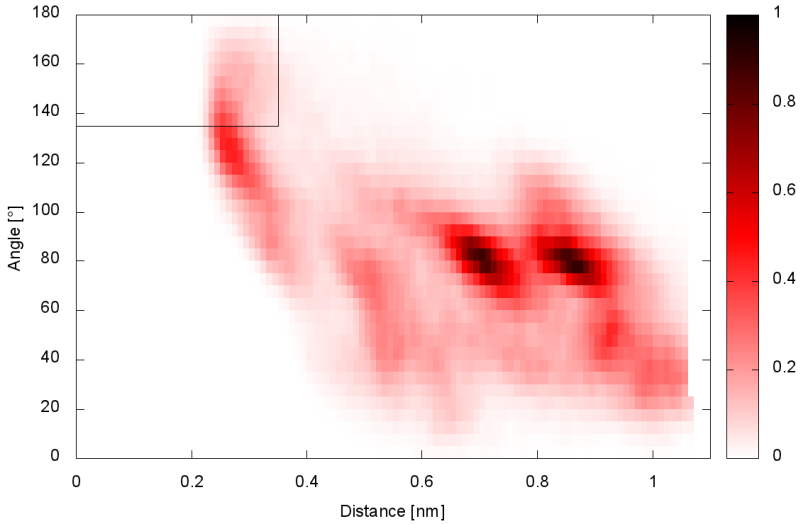


Figure S16: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant FL and 1-ethyl-4-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

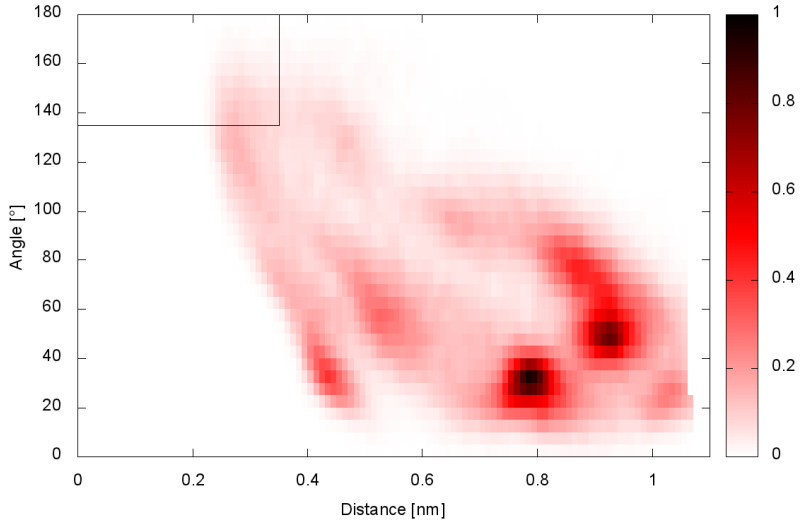


Figure S17: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant FL and 1-ethyl-4-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

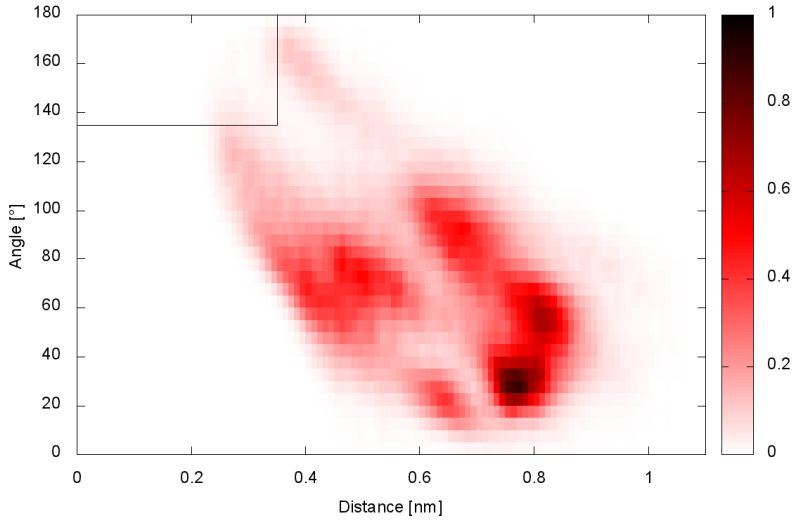


Figure S18: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant NL and 1-ethyl-2-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

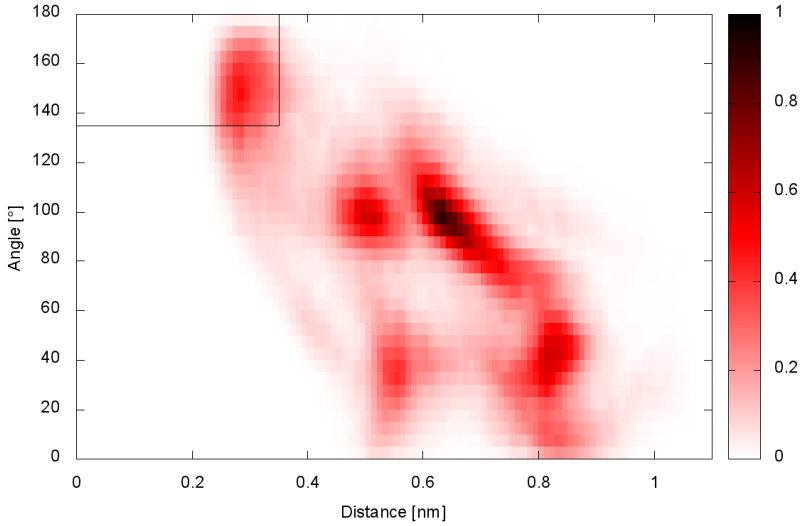


Figure S19: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant NL and 1-ethyl-2-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

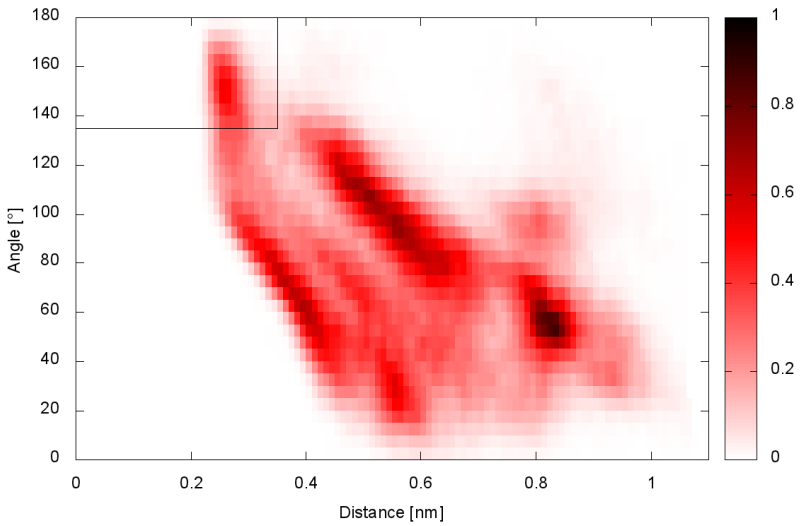


Figure S20: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant NL and 1-ethyl-3-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

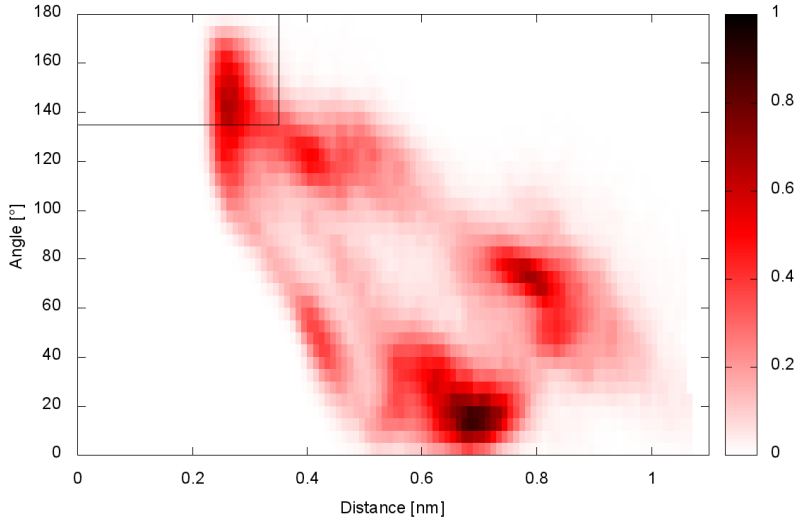


Figure S21: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant NL and 1-ethyl-3-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

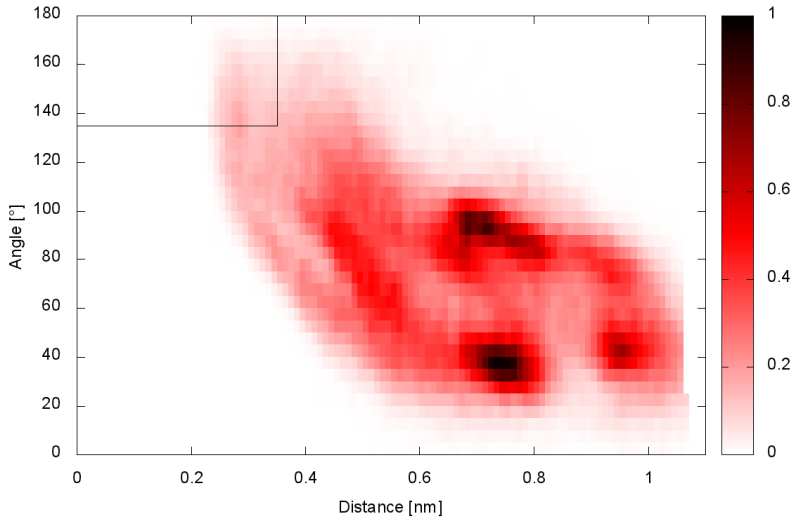


Figure S22: Heat map representing distance dH-O and angle αC-H-O values for pro-R near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant NL and 1-ethyl-4-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

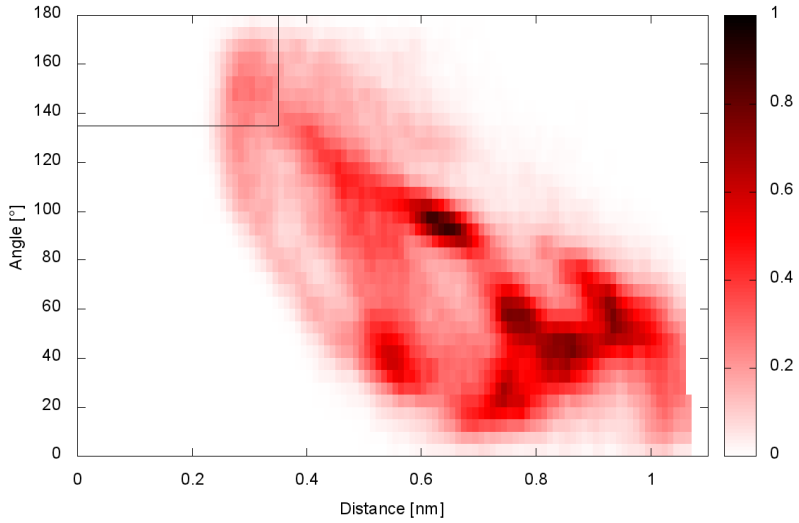


Figure S23: Heat map representing distance dH-O and angle αC-H-O values for pro-S near attack conformations of the substrate in 10 molecular dynamics simulations of CYP101A1 variant NL and 1-ethyl-4-methylbenzene. Heat map was plotted on a grid consisting of 5° × 0.01nm cells. Colors from white through red to black on the heat map represent low to high number of simulation frames representing substrate orientation in the cell of the grid, the plot was normalized.

6.1.1 Molecular dynamics simulations

Molecular dynamics simulations of CYP101A1 (PDB code: 1PHG)[96] variants (Y96F, VF (M184V/T185F), NL (L244N/V247L) and FL (L244F/V247L)) with 1-ethyl-2-methylbenzene, 1-ethyl-3-methylbenzene, 1-ethyl-4-methylbenzene in the binding pocket were performed. The structures of variants were created using mutagenesis tool of the PyMOL program.[211] The substrate molecules were manually placed in the binding pocket using PyMOL. All substrates were placed in the binding pocket in a way that both hydrogens of the stereocenter carbon were between 0.25 nm and 0.5 nm from the compound I ferryl oxygen. The benzyl ring was placed to be pointing into the direction of residues 87 and 96 which are part of the main substrate access channel. Ten molecular dynamics simulations with different starting orientations of the substrate in the binding pocket were performed, to ensure extensive conformational sampling for each variant-substrate complex.

Ten 30 ns simulations per variant-substrate complex were performed using AMBER03 force field[250] in Gromacs version 5.0.4.[251] All water molecules were deleted from the crystal structure. Heme and its cysteine ligand were replaced by a structure of heme compound I covalently bound to cysteine. The previously described field for the cysteine-heme compound I complex was used.[252] Force fields for all substrate molecules (ethylbenzene, 1-ethyl-2-methylbenzene, 1-ethyl-3-methylbenzene, 1-ethyl-4-methylbenzene) were generated based on

the structures derived from PubChem database.[253] The molecules were subsequently subjected to energy minimization in YASARA with AMBER03 force field.[254] The RESP partial charges of the molecules were calculated using R.E.D webserver with RESP-A1B charge model,[255–257] the force field was built with ANTECHAMBER module of AMBER 10,[258] and converted to the GROMACS topology format. Octahedral water box of SPC/E water with periodic boundaries at least 1.2 nm from the protein on all sides was used. Simulations were run at 300 K and 1 bar. Pressure coupling was performed with Parrinello-Rahman barostat.[259] The Nose-Hoover coupling scheme was used to maintain the temperature, with coupling constant of 0.5 ps.[260] Initial velocities were randomly assigned. LINICS algorithm was applied to constrain all bonds containing hydrogen atoms.[251] Eighteen Na+ counter ions were added to maintain the neutral charge of the systems. Long range electrostatic interactions were treated by using the particle-mesh Ewald method.[261] Energy minimization was performed using the steepest descent method with positional restraints on protein heavy atoms and a maximum allowed force of 1000 kJ/mol/nm. A 2 fs time step was used and coordinates were saved every 1000 steps (2 ps). All systems were energy minimized, gradually heated to 300 K over 1 ns and the simulations were continued for another 30 ns. Final 25 ns of the simulations were used for analyses.

The selectivity prediction method is based on previously published work, which allowed authors to predict selectivity of CYP101A1 based on molecular dynamics simulations.[114] The method is based on measurement of dH-O (a distance between the ferryl oxygen and hydrogen of the substrate) and angle αC-H-O (an angle between ferryl oxygen and hydrogen and carbon of the substrate) (Figure 15 on page 52). Pro-R orientations were assigned when αC-H-O for the R-hydrogen was 180±45°, the dH-O shorter than 0.35 nm and shorter than the distance for S-hydrogen. The pro-S orientations were assigned analogously. All snapshots of the MD simulations were classified based on those criteria and assigned to groups of pro-R, pro-S or non-productive orientations. For each variant–substrate complex, the numbers of pro-R and pro-S orientations from ten simulation runs were summed up and percentages were calculated. The percentages of pro-R and pro-S orientations were used as an estimate for the experimental product formation and enantiomeric excess values.

6.2 Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance

Table S1: Equivalence table between numbering of class I and class II CYPs. The table is based on the structural alignment of the class I reference CYP101A1 (pdb:1PHG) and the class II reference CYP102A1 (pdb:1ZOA). Numbers shown in the table are the native position numbers of the class I and class II reference sequences and are therefore the standard numbers for the class I and the class II numbering schemes. Numbers in the same rows are representing structurally corresponding positions. The table allows for the comparison of sequences numbered using the two different numbering schemes (Figure S24C).

Class I	Class II
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
	1
	2
	3
	4
21	5
22	6
23	7
24	8
25	
26	
	9
	10
	11
	12
	13
	14
27	15
28	16
29	17
30	18
31	19
32	
33	
34	
35	
36	
	20
	21
	22
	23

Class I	Class II
37	24
38	25
39	26
40	27
41	28
42	29
43	30
44	31
45	32
46	33
47	34
48	
49	35
50	36
51	37
52	38
53	39
54	40
55	41
56	42
57	43
58	
59	
	44
	45
	46
60	47
61	48
62	49
63	50
64	51
65	52
66	53
67	54
68	55
69	56
70	57
71	58
72	59
73	60
74	61
75	62
76	
77	63
78	64
79	65
80	66
81	67
82	68
83	69

Class I	Class II
84	70
85	71
86	72
87	73
88	74
89	
90	
91	
92	75
93	76
94	
	77
95	78
96	79
	80
	81
	82
97	83
98	84
99	85
100	86
101	87
102	88
103	89
104	90
105	91
106	92
	93
107	94
	95
108	96
109	97
110	98
111	99
112	100
113	101
114	102
115	103
116	104
117	105
118	106
119	107
120	108
121	109
122	110
123	111
124	112
125	113
126	114
127	115

Class I	Class II
128	116
129	117
130	118
131	119
132	120
133	121
134	122
135	123
136	124
137	125
138	126
139	127
140	128
141	129
142	130
143	131
	132
144	133
145	134
	135
	136
146	137
147	138
148	139
149	140
150	141
151	142
152	143
153	144
154	
155	145
156	146
157	147
158	148
159	149
160	150
161	151
162	152
163	153
164	154
	155
165	156
166	157
167	158
168	159
169	160
170	161
171	162
172	163
173	

Class I	Class II
174	
	164
	165
	166
	167
	168
	169
	170
175	171
176	172
177	173
178	174
179	175
180	176
181	177
182	178
183	179
184	180
185	181
186	182
	183
	184
187	185
188	186
189	
	187
	188
	189
	190
	191
	192
	193
	194
	195
	196
	197
	198
	199
	200
190	201
191	202
	203
192	204
193	205
194	206
195	207
196	208
197	209
198	210
199	211

Class I	Class II
200	212
201	213
202	214
203	215
204	216
205	217
206	218
207	219
208	220
209	221
210	222
211	223
212	224
213	225
214	226
215	
	227
	228
	229
216	230
217	231
218	232
219	233
220	234
221	235
222	236
223	237
224	238
225	239
226	240
227	241
228	242
	243
	244
229	245
230	246
231	247
232	248
233	249
234	250
235	251
236	252
237	253
238	254
239	255
240	256
241	257
242	258
243	259
244	260

Class I	Class II
245	261
246	262
247	263
248	264
249	265
250	266
251	267
252	268
253	269
254	270
255	271
256	272
257	273
258	274
259	275
260	276
261	277
262	278
263	279
264	280
265	281
266	282
267	283
268	284
269	285
270	286
271	287
272	288
273	289
274	290
275	291
276	292
277	293
278	294
	295
	296
	297
	298
	299
	300
	301
	302
	303
	304
	305
	306
	307
	308
	309
	310

Class I	Class II
	311
279	312
280	313
281	314
282	315
283	316
284	317
285	318
286	319
287	320
288	321
289	322
290	323
291	324
292	325
293	326
294	327
295	328
	329
296	330
297	331
298	332
299	333
300	334
301	335
302	336
303	337
304	338
305	339
306	340
307	341
308	342
	343
309	344
310	345
311	346
312	347
313	348
314	349
315	350
316	351
317	352
318	353
319	354
320	355
321	356
322	357
323	358
324	359
325	360

Class I	Class II
326	361
327	362
328	363
329	364
330	365
331	366
332	367
333	368
	369
334	370
335	371
336	372
337	373
338	374
339	375
340	376
341	377
342	378
	379
	380
343	381
	382
	383
344	384
345	385
	386
	387
	388
346	389
347	390
348	391
349	392
350	393
351	394
352	395
353	396
354	397
355	398
356	399
357	400
358	401
359	402
360	403
361	404
362	405
363	406
364	407
365	408
366	409
367	410

Class I	Class II
368	411
369	412
370	413
371	414
372	415
373	416
374	417
375	418
376	419
377	420
378	421
379	
380	422
381	423
382	424
383	425
384	426
385	427
386	428
387	429
388	430
389	431
390	432
391	433
392	434
393	435
394	436
395	437
396	438
397	439
398	440
399	441
400	442
401	443
402	444
403	445
404	446
405	447
406	448
407	449
408	450
409	451
410	
411	
412	452
413	453
414	454
	455
	456
	457
	458
	459

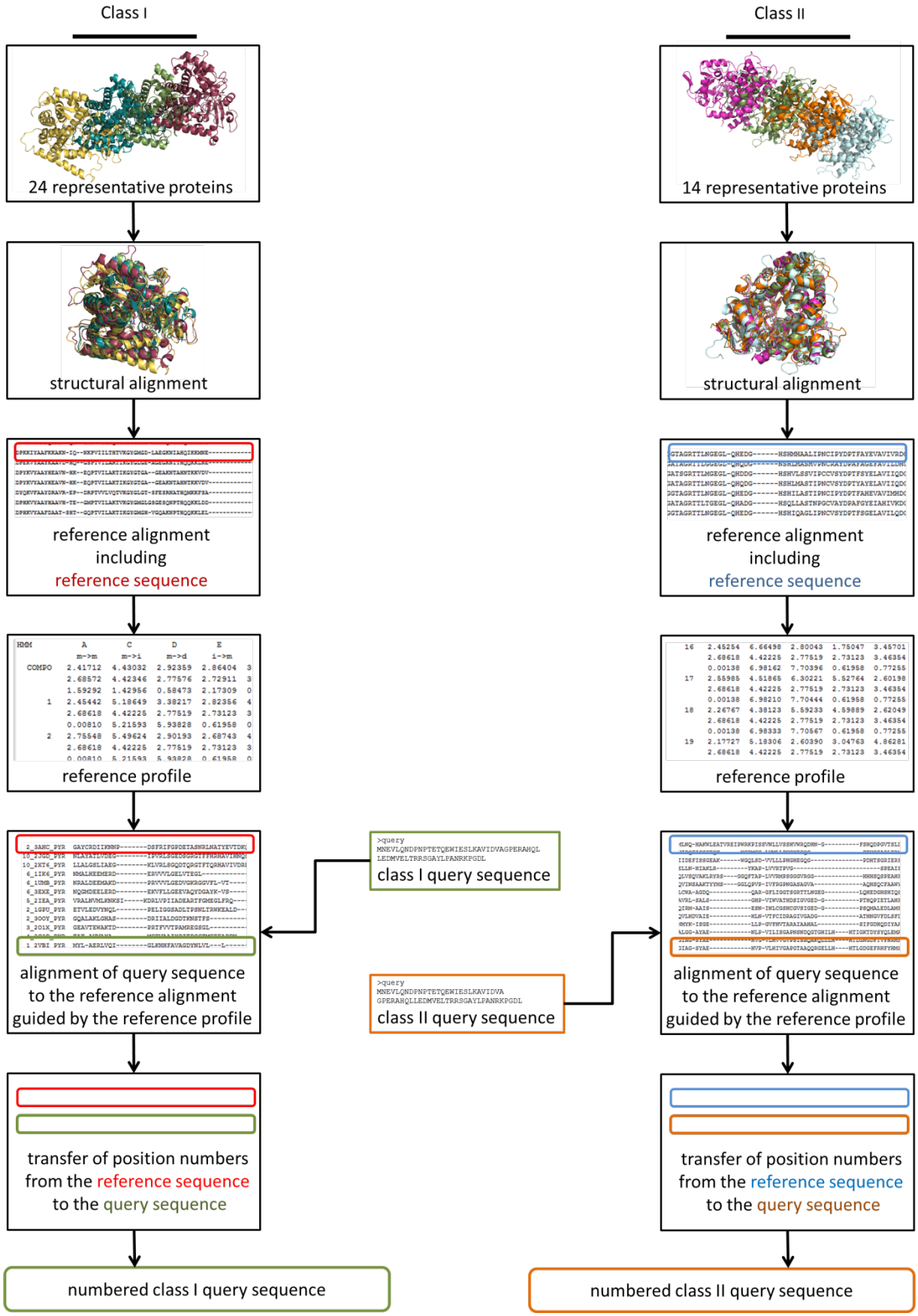
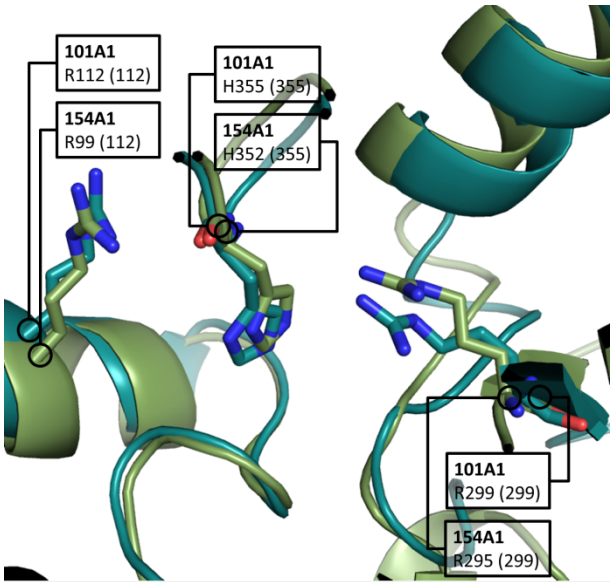
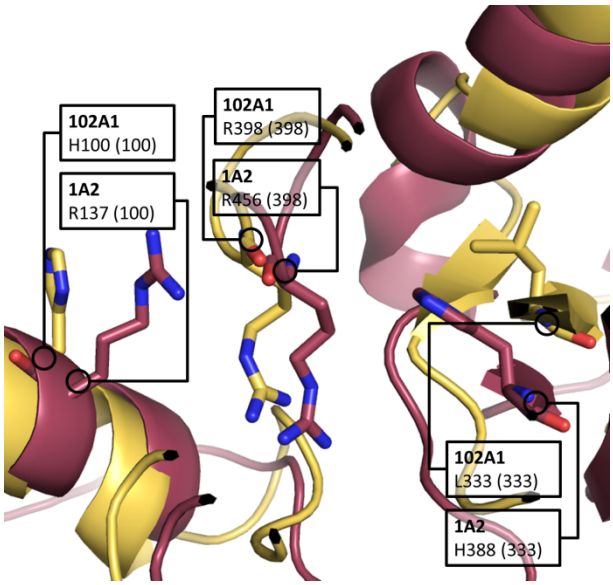


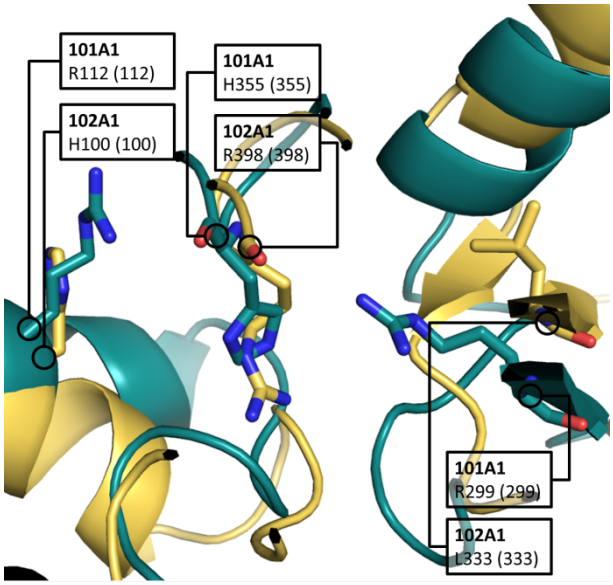
Figure S24: Flowchart of the numbering scheme



A

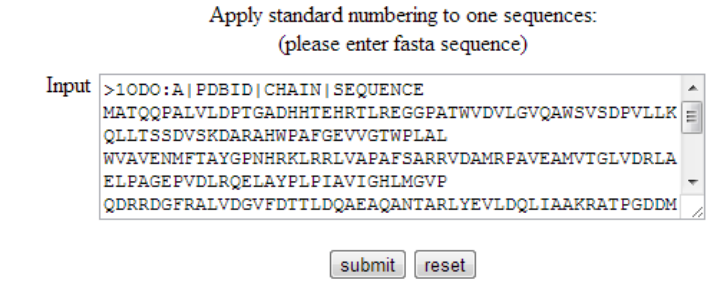


B



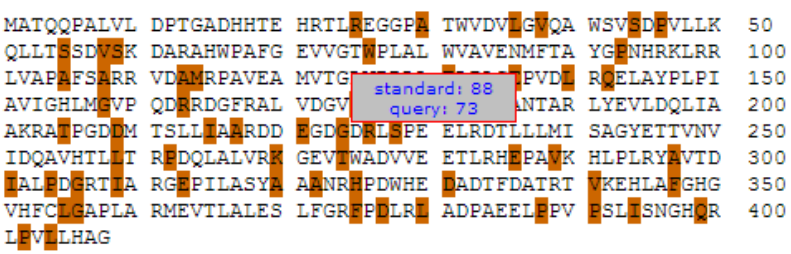
C

Figure S25: Structural superimposition of core amino acids of CPYs 101A1 from *P. putida* (cyan/PDB ID: 1PHG), 102A1 from *B. megaterium* (yellow/PDB ID: 1ZOA), 154A1 from *S. coelicolor* (green/PDB ID: 1ODO) and 1A2 from *H. sapiens* (red/PDB ID: 2HI4). CYP101A1 was chosen as the reference sequence for the class I and CYP102A1 as the reference sequence for the class II numbering scheme. The residue numbers indicate the absolute number of the respective sequence and the numbers in brackets show the standard number of the respective position. A+B: The overlays show structural equivalence of the residues with identical standard numbers in the respective classes I (A) and II (B) and high structural similarity within the classes. C: The structures of the class I and class II references deviate in the superimposition but structural equivalence was found for several positions. Despite deviating standard numbers in a comparison of a class I to a class II CYP, the corresponding positions can be determined using the equivalence table (Table S1).



A

p450 class 1 numbering scheme applied to your query sequence:

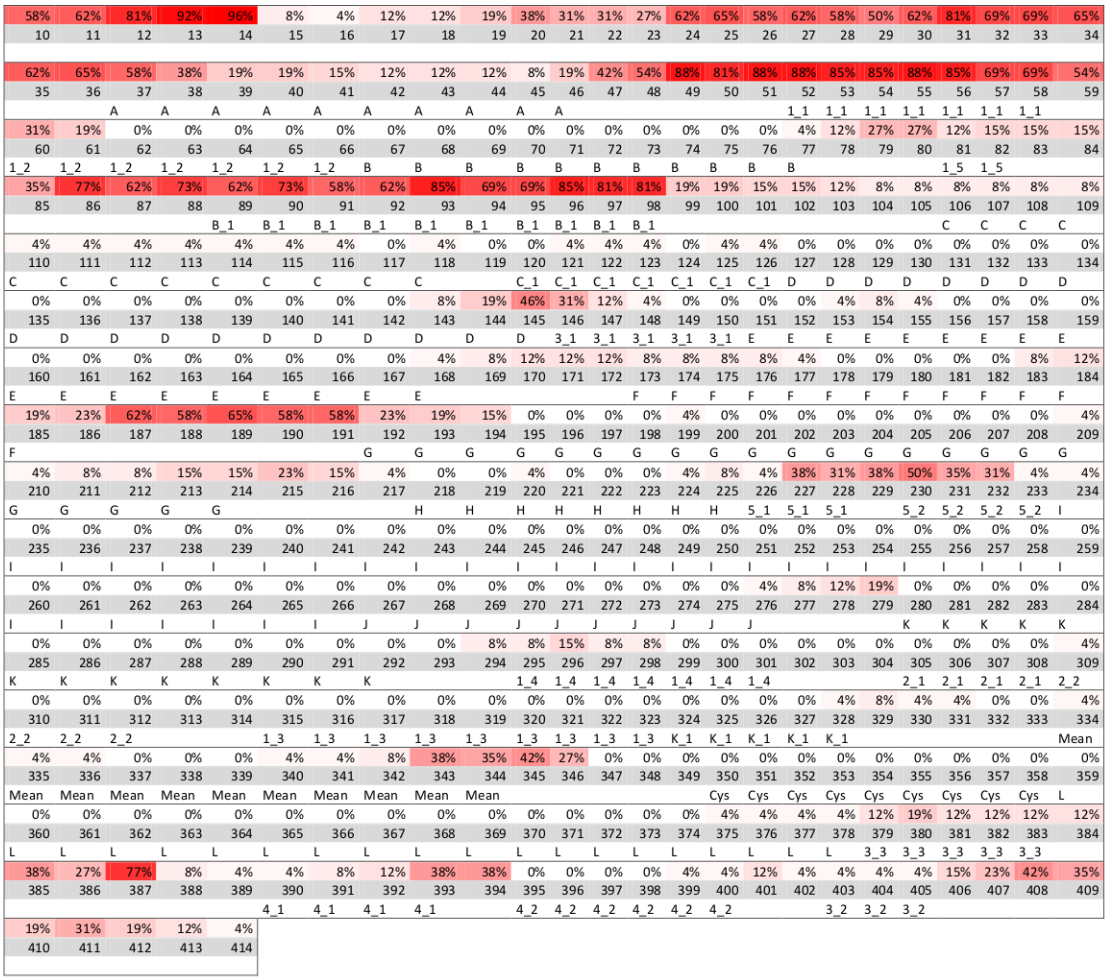


show/hide details

absolute position number	amino acid	standard position number
1	M	3
2	A	4
3	T	5
4	Q	6
5	Q	7

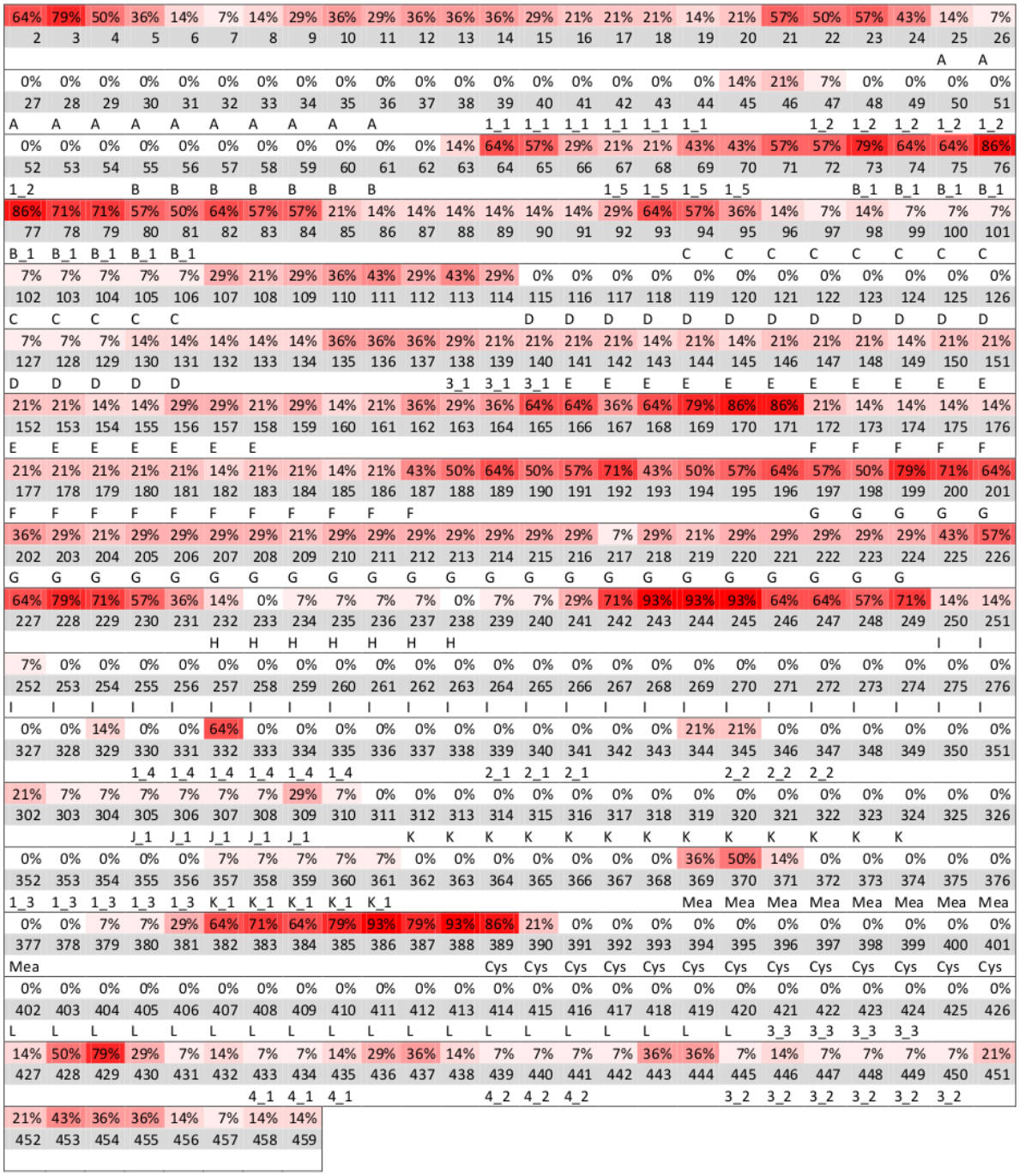
B

Figure S26: Screenshots of the web application of the CYP numbering scheme. A: Submission form for a query sequence in FASTA format. B: In the results view of the numbering scheme, the numbered query sequence is shown with highlighted annotations (boundaries of predicted secondary structure elements). By moving the mouse cursor over distinct amino acids the interface will display information about the standard and absolute position number as well as position specific annotations. In addition, the interface provides a detailed view with a conversion table showing all residues of the query sequence with their absolute and their standard position number.



58%	62%	81%	92%	96%	8%	4%	12%	12%	19%	38%	31%	31%	27%	62%	65%	58%	62%	58%	50%	62%	81%	69%	69%	65%
10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
62%	65%	58%	38%	19%	19%	15%	12%	12%	12%	8%	19%	42%	54%	88%	81%	88%	88%	85%	85%	88%	85%	69%	69%	54%
35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
		A	A	A	A	A	A	A	A	A	A						1_1	1_1	1_1	1_1	1_1	1_1	1_1	
31%	19%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	12%	27%	27%	12%	15%	15%	15%
60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84
1_2	1_2	1_2	1_2	1_2	1_2	1_2	B	B	B	B	B	B	B	B	B	B	B				1_5	1_5		
35%	77%	62%	73%	62%	73%	58%	62%	85%	69%	69%	85%	81%	81%	19%	19%	15%	15%	12%	8%	8%	8%	8%	8%	8%
85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109
				B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1	B_1								C	C	C	C
4%	4%	4%	4%	4%	4%	4%	0%	4%	0%	0%	4%	4%	4%	0%	4%	4%	0%	0%	0%	0%	0%	0%	0%	0%
110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134
C	C	C	C	C	C	C	C	C		C_1	C_1	C_1	C_1	C_1	C_1	C_1	D	D	D	D	D	D	D	D
0%	0%	0%	0%	0%	0%	0%	0%	8%	19%	46%	31%	12%	4%	0%	0%	0%	0%	4%	8%	4%	0%	0%	0%	0%
135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
D	D	D	D	D	D	D	D	D	D	D	3_1	3_1	3_1	3_1	3_1	E	E	E	E	E	E	E	E	E
0%	0%	0%	0%	0%	0%	0%	0%	4%	8%	12%	12%	12%	8%	8%	8%	8%	4%	0%	0%	0%	0%	0%	8%	12%
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184
E	E	E	E	E	E	E	E	E					F	F	F	F	F	F	F	F	F	F	F	F
19%	23%	62%	58%	65%	58%	58%	23%	19%	15%	0%	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%
185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209
F							G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
4%	8%	8%	15%	15%	23%	15%	4%	0%	0%	4%	0%	0%	0%	4%	8%	4%	38%	31%	38%	50%	35%	31%	4%	4%
210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234
G	G	G	G	G				H	H	H	H	H	H	H	H	5_1	5_1	5_1		5_2	5_2	5_2	5_2	I
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259
I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	8%	12%	19%	0%	0%	0%	0%	0%
260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284
I	I	I	I	I	I	I	J	J	J	J	J	J	J	J	J	J				K	K	K	K	K
0%	0%	0%	0%	0%	0%	0%	0%	0%	8%	8%	15%	8%	8%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%
285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309
K	K	K	K	K	K	K	K			1_4	1_4	1_4	1_4	1_4	1_4	1_4				2_1	2_1	2_1	2_1	2_2
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	8%	4%	4%	0%	0%	4%
310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334
2_2	2_2	2_2			1_3	1_3	1_3	1_3	1_3	1_3	1_3	1_3	1_3	K_1	K_1	K_1	K_1	K_1								Mean
4%	4%	0%	0%	0%	4%	4%	8%	38%	35%	42%	27%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359
Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean						Cys	Cys	Cys	Cys	Cys	Cys	Cys	Cys	Cys	L
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	4%	4%	4%	12%	19%	12%	12%	12%	12%
360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384
L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	3_3	3_3	3_3	3_3	3_3	
38%	27%	77%	8%	4%	4%	8%	12%	38%	38%	0%	0%	0%	0%	4%	4%	12%	4%	4%	4%	4%	15%	23%	42%	35%
385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409
					4_1	4_1	4_1	4_1		4_2	4_2	4_2	4_2	4_2	4_2			3_2	3_2	3_2				
19%	31%	19%	12%	4%																				
410	411	412	413	414																				

Figure S27A: Numbering scheme validation for class I by a comparison of the numbering scheme based alignments and structural alignments for class I. Each row consists of three sub rows: the first sub row represents the percentage of different alignment column obtained using two different alignment methods. Pairwise alignments of 26 CYPs and the class I reference CYP101A1 were performed using the numbering scheme and STAMP. Subsequently, the results from both methods were compared by checking the identity for each column of the resulting alignments. The average difference of the alignment columns between the two alignments is given as percentages. The second sub row shows numbering scheme position and the third sub row shows the structural element names (Helices: A-L, Beta strands: 1_1 - 5_2, meander loop: Mean and cysteine pocket: Cys). Red gradient in the first sub row reflects the percentages from white – 0% to bright red 100%.



64%	79%	50%	36%	14%	7%	14%	29%	36%	29%	36%	36%	36%	29%	21%	21%	21%	14%	21%	57%	50%	57%	43%	14%	7%
2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
																							A	A
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	14%	21%	7%	0%	0%	0%	0%
27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51
A	A	A	A	A	A	A	A	A	A			1_1	1_1	1_1	1_1	1_1	1_1			1_2	1_2	1_2	1_2	1_2
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	14%	64%	57%	29%	21%	21%	43%	43%	57%	57%	79%	64%	64%	86%
52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
1_2			B	B	B	B	B	B	B						1_5	1_5	1_5	1_5				B_1	B_1	B_1	B_1
86%	71%	71%	57%	50%	64%	57%	57%	21%	14%	14%	14%	14%	14%	14%	29%	64%	57%	36%	14%	7%	14%	7%	7%	7%
77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101
B_1	B_1	B_1	B_1	B_1													C	C	C	C	C	C	C	C
7%	7%	7%	7%	7%	29%	21%	29%	36%	43%	29%	43%	29%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126
C	C	C	C	C									D	D	D	D	D	D	D	D	D	D	D	D
7%	7%	7%	14%	14%	14%	14%	14%	36%	36%	36%	29%	21%	21%	21%	21%	14%	21%	14%	21%	21%	21%	14%	21%	21%
127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151
D	D	D	D	D							3_1	3_1	3_1	E	E	E	E	E	E	E	E	E	E	E
21%	21%	14%	14%	29%	29%	21%	29%	14%	21%	36%	29%	36%	64%	64%	36%	64%	79%	86%	86%	21%	14%	14%	14%	14%
152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176
E	E	E	E	E	E	E														F	F	F	F	F
21%	21%	21%	21%	21%	14%	21%	21%	14%	21%	43%	50%	64%	50%	57%	71%	43%	50%	57%	64%	57%	50%	79%	71%	64%
177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201
F	F	F	F	F	F	F	F	F	F	F										G	G	G	G	G
36%	29%	21%	29%	29%	29%	29%	21%	29%	29%	29%	29%	29%	29%	29%	7%	29%	21%	29%	29%	29%	29%	29%	43%	57%
202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226
G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G		
64%	79%	71%	57%	36%	14%	0%	7%	7%	7%	7%	0%	7%	7%	29%	71%	93%	93%	93%	64%	64%	64%	57%	71%	14%	14%
227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251
					H	H	H	H	H	H	H												I	I
7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276
I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
0%	0%	14%	0%	0%	64%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	21%	21%	0%	0%	0%	0%	0%	0%
327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351
			1_4	1_4	1_4	1_4	1_4	1_4				2_1	2_1	2_1				2_2	2_2	2_2				
21%	7%	7%	7%	7%	7%	7%	29%	7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326
			J_1	J_1	J_1	J_1	J_1			K	K	K	K	K	K	K	K	K	K	K	K	K		
0%	0%	0%	0%	0%	7%	7%	7%	7%	7%	0%	0%	0%	0%	0%	0%	0%	36%	50%	14%	0%	0%	0%	0%	0%
352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376
1_3	1_3	1_3	1_3	1_3	K_1	K_1	K_1	K_1	K_1								Mea	Mea	Mea	Mea	Mea	Mea	Mea	Mea
0%	0%	7%	7%	29%	64%	71%	64%	79%	93%	79%	93%	86%	21%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401
Mea												Cys	Cys	Cys	Cys	Cys	Cys	Cys	Cys	Cys	Cys	Cys	Cys	Cys
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426
L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	3_3	3_3	3_3	3_3		
14%	50%	79%	29%	7%	14%	7%	7%	14%	29%	36%	14%	7%	7%	7%	7%	36%	36%	7%	14%	7%	7%	7%	7%	21%
427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451
						4_1	4_1	4_1				4_2	4_2	4_2				3_2	3_2	3_2	3_2	3_2	3_2	
21%	43%	36%	36%	14%	7%	14%	14%																	
452	453	454	455	456	457	458	459																	

Figure S27B: Numbering scheme validation for class II by a comparison of the numbering scheme based alignments and structural alignments for class II. Each row consists of three sub rows: the first sub row represents the percentage of different alignment column obtained using two different alignment methods. Pairwise alignments of 14 CYPs and the class II reference CYP102A1 were performed using the numbering scheme and STAMP. Subsequently, the results from both methods were compared by checking the identity for each column of the resulting alignments. The average difference of the alignment columns between the two alignments is given as percentages. The second sub row shows numbering scheme position and the third sub row shows the structural element names (Helices: A-L, Beta strands: 1_1 - 5_2, meander loop: Mea and cysteine pocket: Cys). Red gradient in the first sub row reflects the percentages from white – 0% to bright red 100%.

6.3 Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining

Due to its size (over 150 pages) Table S2 is available in the complete supporting information on the publishers' website.

Table S3: Frequency of CYPs with literature information for different positions. Conserved positions were colored: positions with single conserved amino acid are red, and positions where certain properties (aromatic, charged or hydrogen binding) were conserved are colored yellow.

Standard position	Number of CYPs with literature information about the position	Functionally significant region	Standard position	Number of CYPs with literature information about the position	Functionally significant region
1	*6*		*46*	*5*	
2	3		47	19	
3	4		48	8	
3.5	1		49	7	
4	1		50	4	
5	4		51	11	
6	2		52	3	
7	2		53	4	
8	2		54	4	
9	2		55	5	
9.1	2		56	1	
10	1		57	1	
11	5		58	3	
11.1	6		59	8	
12	1		60	4	
13	2		61	3	
14	2		62	8	
15	3		62.1	5	
16	2		62.2	5	
17	3		63	6	
17.2	1		64	4	
18	0		65	6	
19	4		65.1	1	
20	4		65.3	1	
21	3		66	3	
22	8		67	4	
23	2		68	4	
24	2		69	17	SRS1
25	8		70	21	SRS1
26	8		71	22	SRS1
27	3		72	19	SRS1
28	3		73	34	SRS1
29	7		74	23	SRS1
30	2		75	28	SRS1
31	1		75.1	1	SRS1
32	5		76	16	SRS1
33	5		77	15	SRS1
34	4		78	15	SRS1
35	3		78.1	2	SRS1
36	4		79	26	SRS1
37	2		80	5	SRS1
38	3		81	20	SRS1
39	4		81.2	1	SRS1
40	3		82	23	SRS1
41	3		83	10	SRS1
42	6		84	14	SRS1
43	9		85	25	SRS1
44	9		86	28	SRS1
45	7		86.1	1	SRS1
45.1	1		86.6	4	SRS1

Standard position	Number of CYPs with literature information about the position	Functionally significant region	Standard position	Number of CYPs with literature information about the position	Functionally significant region
87	63	SRS1	136	6	
88	42	SRS1	137	5	
89	13	SRS1	138	7	
90	10	SRS1	139	5	
90.1	1	SRS1	140	7	
91	5	SRS1	141	3	
92	3	SRS1	142	11	
93	0		143	5	
94	8		144	4	
95	3		145	3	
96	25		146	11	
97	16		147	2	
98	6		148	5	
99	8		149	4	
100	28		150	8	
101	12		151	7	
102	8		152	7	
103	13		153	5	
104	11		154	6	
105	3		155	4	
106	8		156	4	
107	10		157	4	
108	11		158	6	
109	9		159	6	
110	8		160	4	
111	2		161	5	
112	5		161.9	1	
112.1	1		162	7	
112.5	3		163	6	
112.6	1		164	0	
113	10		165	4	
114	2		166	6	
115	1		167	2	
116	4		168	4	
117	7		169	4	
118	5		170	7	
119	1		171	2	
120	4		172	7	
121	3		173	10	
122	7		174	8	
123	0		175	7	
124	7		176	12	
125	2		177	26	
126	3		178	18	
127	3		179	7	
128	7		180	30	
129	3		180.2	1	
130	3		181	47	SRS2
131	5		182	17	SRS2
132	6		183	11	SRS2
133	3		184	32	SRS2
133.3	1		185	25	SRS2
134	5		185.1	7	SRS2
135	2		185.2	3	SRS2

Standard position	Number of CYPs with literature information about the position	Functionally significant region	Standard position	Number of CYPs with literature information about the position	Functionally significant region
185.3	2	SRS2	226.1	1	
185.4	2	SRS2	226.2	1	
186	19	SRS2	226.5	1	
187	12	SRS2	226.6	5	
188	10	SRS2	227	3	
188.1	1		228	2	
189	12		229	7	
190	7		230	6	
191	10		231	14	
191.1	9		232	7	
191.2	1		233	6	
191.3	4		234	4	
191.4	1		235	4	
191.5	2		236	3	
192	12		237	5	
192.1	3		238	3	
193	5		239	5	
194	7		240	3	
194.1	1		241	3	
195	5		242	6	
196	5		242.1	6	
197	5		243	2	
198	4		244	4	
199	3		245	6	
200	1	SRS3	245.1	1	
201	6	SRS3	245.8	5	
201.16	1	SRS3	246	4	
202	4	SRS3	247	4	
203	13	SRS3	248	3	
204	7	SRS3	249	4	
205	16	SRS3	250	4	
206	16	SRS3	251	7	
207	7	SRS3	252	9	
208	13	SRS3	253	10	SRS4
209	24		254	5	SRS4
210	8		255	12	SRS4
211	6		256	20	SRS4
212	13		257	7	SRS4
213	15		258	9	SRS4
214	3		259	19	SRS4
215	5		260	42	SRS4
216	7		261	18	SRS4
217	1		262	8	SRS4
218	5		263	44	SRS4
219	5		264	57	SRS4
220	5		265	22	SRS4
221	11		266	24	SRS4
222	2		267	46	SRS4
223	5		268	60	SRS4
224	6		269	16	SRS4
225	6		270	12	SRS4
226	7		271	14	SRS4

Standard position	Number of CYPs with literature information about the position	Functionally significant region	Standard position	Number of CYPs with literature information about the position	Functionally significant region
272	7		321	2	
273	4		322	5	
274	8		323	15	
275	10		324	8	
276	3		325	15	SRS5
277	3		326	11	SRS5
278	4		327	33	SRS5
279	1		328	57	SRS5
280	8		329	13	SRS5
281	3		330	39	SRS5
282	3		331	35	SRS5
283	5		331.1	25	SRS5
284	0		332	25	SRS5
285	7		333	37	SRS5
286	2		334	8	SRS5
287	3		335	5	SRS5
288	6		336	6	
289	7		337	5	
290	4		338	5	
291	5		339	5	
292	3		340	5	
293	13		341	3	
294	7		342	4	
295	4		343	2	
296	7		344	0	
297	2		345	7	
298	4		346	5	
299	2		347	0	
300	3		348	5	
300.1	3		349	6	
301	6		350	0	
302	3		350.1	1	
303	5		351	7	
304	2		352	7	
305	2		353	4	
306	3		354	16	
307	3		355	4	
308	0		356	6	
309	2		357	10	
310	6		358	5	
310.1	1		359	2	
310.2	5		360	5	
311	2		361	5	
312	8		362	5	
313	2		363	5	
314	4		364	0	
315	4		365	3	
316	4		366	6	
317	0		367	5	
318	2		368	3	
319	11		369	2	
320	9		370	3	

Standard position	Number of CYPs with literature information about the position	Functionally significant region	Standard position	Number of CYPs with literature information about the position	Functionally significant region
371	9		417	5	
372	2		418	3	
373	5		419	3	
374	6		420	4	
375	3		421	0	
376	4		422	1	
377	6		423	4	
378	11		424	0	
379	7		425	5	
380	6		425.3	1	
381	0		426	3	
382	2		427	0	
382.1	2		428	2	
382.2	4		428.1	1	
382.3	1		428.3	2	
382.4	7		429	2	
383	4		430	8	
384	6		431	6	
385	6		432	3	
386	4		433	7	
386.1	1		433.1	4	
386.2	1		434	12	SRS6
387	4		435	20	SRS6
388	7		436	17	SRS6
389	9	Cysteine pocket	437	51	SRS6
390	5	Cysteine pocket	437.1	1	SRS6
391	8	Cysteine pocket	437.3	1	SRS6
392	11	Cysteine pocket	437.4	2	SRS6
393	13	Cysteine pocket	438	45	SRS6
394	9	Cysteine pocket	439	9	SRS6
395	6	Cysteine pocket	440	16	SRS6
396	9	Cysteine pocket	441	8	SRS6
397	11	Cysteine pocket	442	8	
398	35	Cysteine pocket	443	7	
399	12	Cysteine pocket	443.1	1	
400	50	Cysteine pocket	444	5	
401	12	Cysteine pocket	445	7	
402	12	Cysteine pocket	446	1	
403	14		447	7	
404	12		448	3	
405	10		449	4	
406	6		450	4	
407	14		451	6	
408	4		452	1	
409	7		453	3	
410	0		454	11	
411	4		455	3	
412	0		456	0	
413	8		457	0	
414	4		458	2	
415	6				
416	3				

6.4 Redox partner interaction sites in cytochrome P450 monooxygenases: in silico analysis and experimental validation

Table S4: Primer for the QuikChange modification of the CYP153A-CPR gene. Modified region in bold.

Mutation	Sequenz	T_m [°C]
CYP153A-CPR_L115K	CGGTTTGAAGACATC **AAA** TTCGTGGATAAGAGTC GACTCTTATCCACGAA **TTT** GATGTCTTCAAACCG	70.9*
CYP153A-CPR_S120D	CTGTTCGTGGATAAG **GAT** CACGACCTGTTTTCCG CGGAAAACAGGCGTG **ATC** CTTATCCACGAACAG	73.8*
CYP153A-CPR_D153K	GATCCGCCGAAACAC **AAA** GTGCAGCGCAGCTCG CGAGCTGCGCTGCAC **TTT** GTGTTTCGGCGGATC	79.0*
CYP153A-CPR_K166Q	GGAGTAGTGGCACCG **CAA** AACCTGAAGGAGATGG CCATCTCCTTCAGGTT **TTG** CGGTGCCACTACTCC	76.2*
CYP153A-CPR_R422Q	CGTTGCATGGGCAAC **CAG** CTGGCTGAACTGCAAC GTTGCAGTTCAGCCAG **CTG** GTTGCCCATGCAACG	76,2*
CYP153A-CPR_E425L	GGCAACCGTCTGGCT **CTG** CTGCAACTGCGCATC GATGCGCAGTTGCAG **CAG** AGCCAGACGGTTGCC	77.4*

* estimated melting temperature OligoCalc http://www.basic.northwestern.edu/biotools/oligocalc.html)

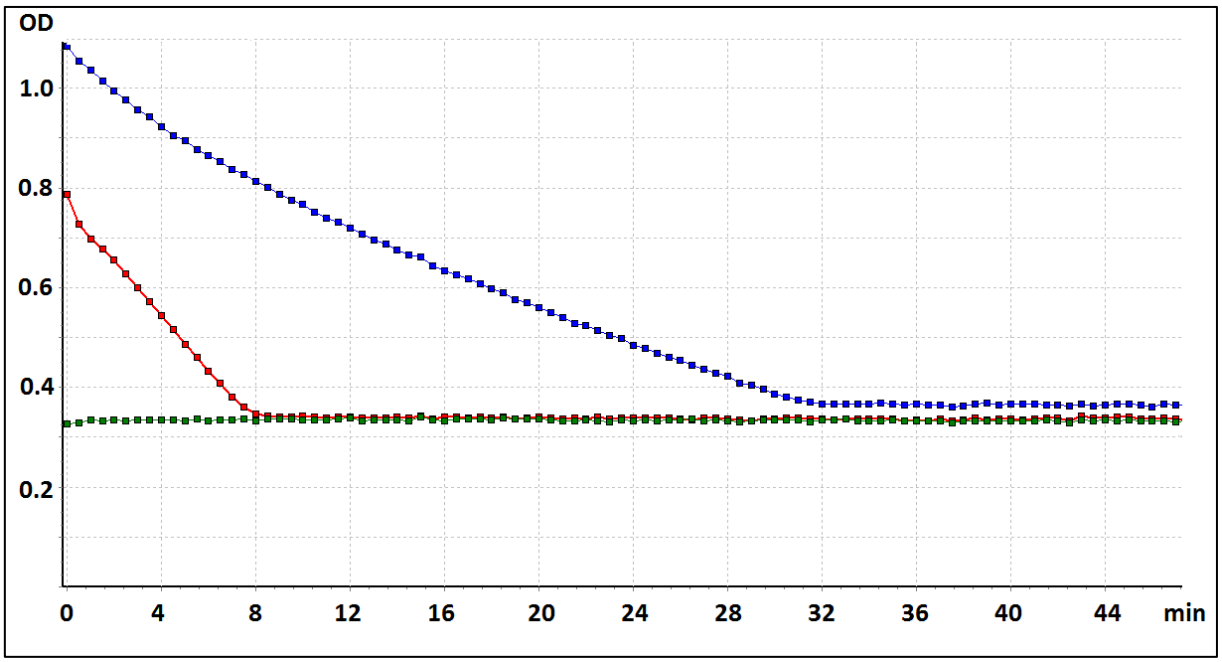


Figure S28: NADPH consumption of CYP153AM.aq.-CPRBM3 wild type (red) and its negative controls without substrate (blue) and without NADPH (green).

Calculation of NADPH-consumption by P450

The slope of the NADPH depletion of the different variants was measured in presence and absence of the model substrate dodecanoic acid. The NADPH depletion in absence of dodecanoic acid was used as a background signal. The difference between the NADPH concentration without substrate and with substrate corresponds to the NADPH consumption of the P450 enzyme. Reactions were stopped after no changes in signal were detected and extracted for the analysis of product formation via GC-FID.

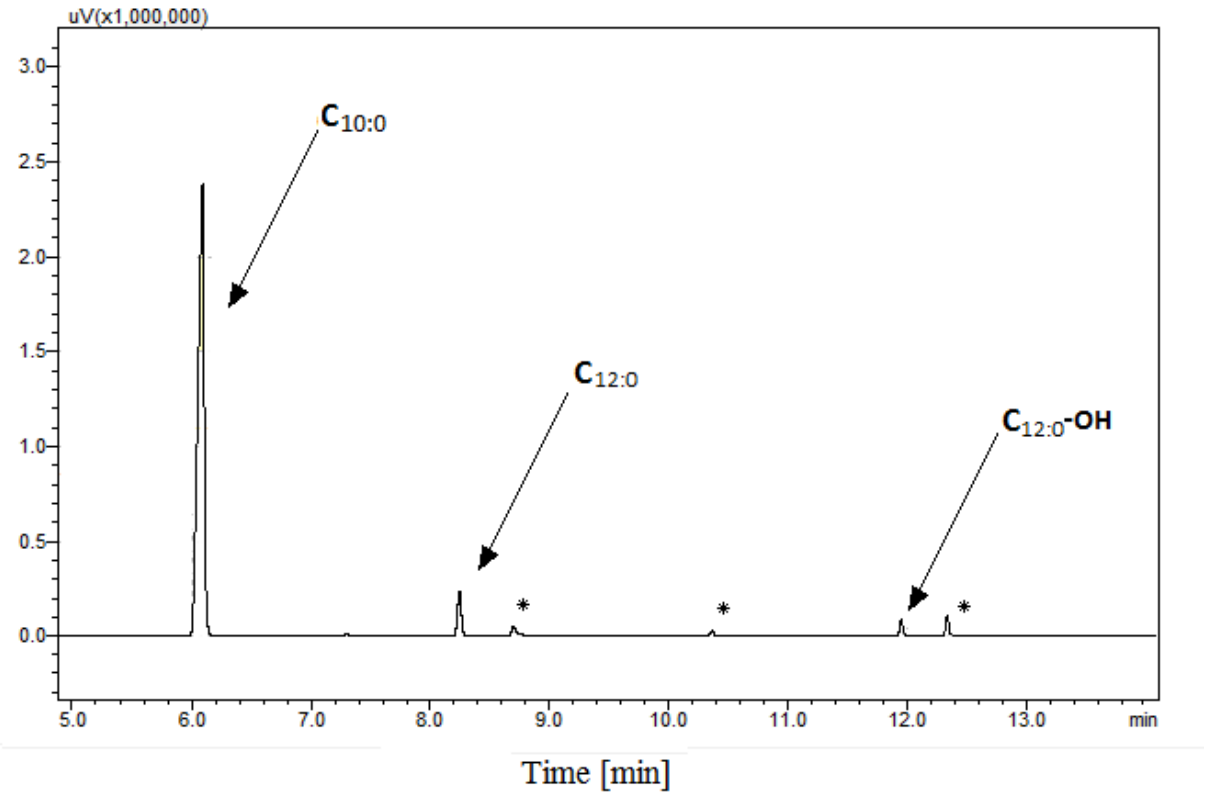


Figure S29: Gas chromatogram for CYP153A-CPR wildtype catalysed reaction with dodecanoic acid. The substrate and formed products were measured as TMS derivatives. Abbreviations: C10:0, decanoic acid as internal standard; C12:0, dodecanoic acid; C12:0-OH ,12-hydroxydodecanoic acid; * impurity.

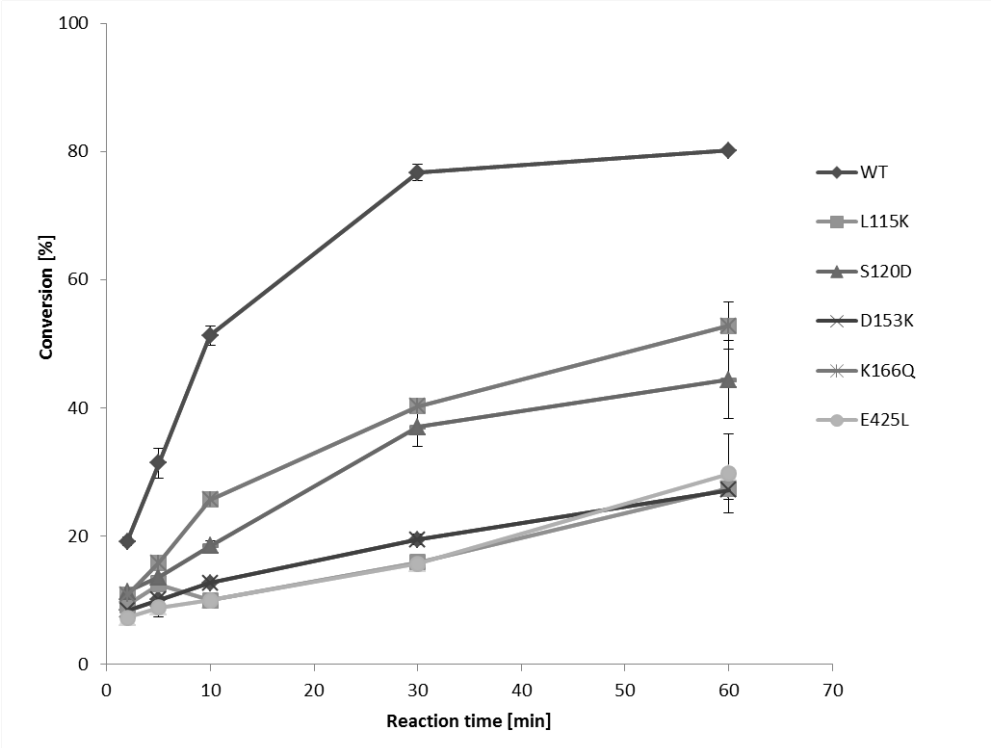


Figure S30: Initial reaction rates of wild type CYP153A6 and RPIS variants.

7. Bibliography

1. Brodie, B. B. *et al.* Detoxication of drugs and other foreign compounds by liver microsomes. *Science* **121,** 603–4 (1955).

2. Axelrod, J. The enzymatic demethylation of ephedrine. *J. Pharmacol. Exp. Ther.* **114,** 430–8 (1955).

3. Garfinkel, D. Studies on pig liver microsomes. I. Enzymic and pigment composition of different microsomal fractions. *Arch. Biochem. Biophys.* **77,** 493–509 (1958).

4. Klingenberg, M. Pigments of rat liver microsomes. *Arch. Biochem. Biophys.* **75,** 376–86 (1958).

5. Omura, T. & Sato, R. The Carbon Monoxide-binding Pigment of Liver Microsomes: I. Evidence for its hemoprotein nature. *J. Biol. Chem.* **239,** 2370–8 (1964).

6. Omura, T. & Sato, R. The Carbon Monoxide-binding Pigment of Liver Microsomes: II Solubilization, purification, and properties. *J. Biol. Chem.* **239,** 2379–85 (1964).

7. Hashimoto, Y., Yamano, T. & Mason, H. S. An electron spin resonance study of microsomal electron transport. *J. Biol. Chem.* **237,** 3843–4 (1962).

8. Werck-Reichhart, D. & Feyereisen, R. Cytochromes P450: a success story. *Genome Biol.* **1,** REVIEWS3003 (2000).

9. Sotaniemi, E. A., Arranto, A. J., Pelkonen, O. & Pasanen, M. Age and cytochrome P450-linked drug metabolism in humans: an analysis of 226 subjects with equal histopathologic conditions. *Clin. Pharmacol. Ther.* **61,** 331–9 (1997).

10. Urlacher, V. B. & Eiben, S. Cytochrome P450 monooxygenases: perspectives for synthetic application. *Trends Biotechnol.* **24,** 324–30 (2006).

11. Bernhardt, R. Cytochromes P450 as versatile biocatalysts. *J. Biotechnol.* **124,** 128–45 (2006).

12. Nelson, D. R. *et al.* The P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers, Early Trivial Names of Enzymes, and Nomenclature. *DNA Cell Biol.* **12,** 1–51 (1993).

13. Nebert, D. W. *et al.* The P450 superfamily: updated listing of all genes and recommended nomenclature for the chromosomal loci. *DNA* **8,** 1–13

14. Nebert, D. W. *et al.* The P450 gene superfamily: recommended nomenclature. *DNA* **6,** 1–11 (1987).

15. Nelson, D. R. The cytochrome p450 homepage. *Hum. Genomics* **4,** 59–65 (2009).

16. Nelson, D. R. Mining databases for cytochrome P450 genes. *Methods Enzymol.* **357,** 3–15 (2002).

17. Katagiri, M., Ganguli, B. N. & Gunsalus, I. C. A soluble cytochrome P-450 functional in methylene hydroxylation. *J. Biol. Chem.* **243,** 3543–6 (1968).

18. Schlichting, I. The Catalytic Pathway of Cytochrome P450cam at Atomic Resolution. *Science (80-.).* **287,** 1615–1622 (2000).

19. Ortiz de Montellano, P. R. & De Voss, J. J. Oxidizing species in the mechanism of cytochrome P450. *Nat. Prod. Rep.* **19,** 477–93 (2002).

20. Denisov, I. G., Makris, T. M., Sligar, S. G. & Schlichting, I. Structure and chemistry of cytochrome P450. *Chem. Rev.* **105,** 2253–77 (2005).

21. Coulter, E. D. Uncoupling Oxygen Transfer and Electron Transfer in the Oxygenation of Camphor Analogues by Cytochrome P450-CAM. *J. Biol. Chem.* **270,** 28042–28048 (1995).

22. Vidal-Limón, A., Águila, S., Ayala, M., Batista, C. V & Vazquez-Duhalt, R. Peroxidase activity stabilization of cytochrome P450(BM3) by rational analysis of intramolecular electron transfer. *J. Inorg. Biochem.* **122,** 18–26 (2013).

23. Cornelissen, S., Liu, S., Deshmukh, A. T., Schmid, A. & Bühler, B. Cell physiology rather than enzyme kinetics can determine the efficiency of cytochrome P450-catalyzed C-H-oxyfunctionalization. *J. Ind. Microbiol. Biotechnol.* **38,** 1359–70 (2011).

24. Juchau, M. R. Substrate specificities and functions of the P450 cytochromes. *Life Sci.* **47,** 2385–94 (1990).

25. Isin, E. M. & Guengerich, F. P. Complex reactions catalyzed by cytochrome P450 enzymes. *Biochim. Biophys. Acta* **1770,** 314–29 (2007).

26. Hannemann, F., Bichet, A., Ewen, K. M. & Bernhardt, R. Cytochrome P450 systems--biological variations of electron transport chains. *Biochim. Biophys. Acta* **1770,** 330–44 (2007).

27. Wright, R. L., Harris, K., Solow, B., White, R. H. & Kennelly, P. J. Cloning of a potential cytochrome P450 from the Archaeon Sulfolobus solfataricus. *FEBS Lett.* **384,** 235–239 (1996).

28. Hlavica, P. Assembly of non-natural electron transfer conduits in the cytochrome P450 system: a critical assessment and update of artificial redox constructs amenable to exploitation in biotechnological areas. *Biotechnol. Adv.* **27,** 103–21 (2009).

29. Nodate, M., Kubota, M. & Misawa, N. Functional expression system for cytochrome P450 genes using the reductase domain of self-sufficient P450RhF from Rhodococcus sp. NCIMB 9784. *Appl. Microbiol. Biotechnol.* **71,** 455–62 (2006).

30. Honda Malca, S. *et al.* Bacterial CYP153A monooxygenases for the synthesis of omega-hydroxylated fatty acids. *Chem. Commun. (Camb).* **48,** 5115–7 (2012).

31. Gricman, Ł., Vogel, C. & Pleiss, J. Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance. *Proteins* **82,** 491–504 (2014).

32. Vogel, C. & Pleiss, J. The modular structure of ThDP-dependent enzymes. *Proteins* **82,** 2523–37 (2014).

33. Racolta, S., Juhl, P. B., Sirim, D. & Pleiss, J. The triterpene cyclase protein family: a systematic analysis. *Proteins* **80,** 2009–19 (2012).

34. Truan, G. & Peterson, J. A. Thr268 in substrate binding and catalysis in P450BM-3. *Arch. Biochem. Biophys.* **349,** 53–64 (1998).

35. Clark, J. P. *et al.* The role of Thr268 and Phe393 in cytochrome P450 BM3. *J Inorg Biochem* **100,** 1075–90 (2006).

36. Altarsha, M., Benighaus, T., Kumar, D. & Thiel, W. How is the reactivity of cytochrome P450cam affected by Thr252X mutation? A QM/MM study for X = serine, valine, alanine, glycine. *J Am Chem Soc* **131,** 4755–63 (2009).

37. Rupasinghe, S. *et al.* The cytochrome P450 gene family CYP157 does not contain EXXR in the K-helix reducing the absolute conserved P450 residues to a single cysteine. *FEBS Lett.* **580,** 6338–42 (2006).

38. Poulos, T. L., Finzel, B. C. & Howard, A. J. High-resolution crystal structure of cytochrome P450cam. *J. Mol. Biol.* **195,** 687–700 (1987).

39. Hasemann, C. A., Kurumbail, R. G., Boddupalli, S. S., Peterson, J. A. & Deisenhofer, J. Structure and function of cytochromes P450:a comparative analysis of three crystal structures. *Structure* **3,** 41–62 (1995).

40. Gotoh, O. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J. Biol. Chem.* **267,** 83–90 (1992).

41. Lee, D. S. . N. P. . H. M. . R. C. S. Structural insights into the evolutionary paths of oxylipin biosynthetic enzymes. *Nature* **455,** 363–368 (2008).

42. Li, L. . C. Z. . P. Z. . F. Z. Q. . W. X. Modes of heme binding and substrate access for cytochrome P450 CYP74A revealed by crystal structures of allene oxide synthase. *Proc.Natl.Acad.Sci.Usa* **105,** 13883–13888 (2008).

43. Sevrioukova, I. F., Li, H., Zhang, H., Peterson, J. A. & Poulos, T. L. Structure of a cytochrome P450-redox partner electron-transfer complex. *Proc.Natl.Acad.Sci.USA* **96,** 1863–1868 (1999).

44. Strushkevich, N. . M. F. . C. T. . G. I. . U. S. . P. H. W. Structural basis for pregnenolone biosynthesis by the mitochondrial monooxygenase system. *Proc.Natl.Acad.Sci.USA* **108,** 10139–10143 (2011).

45. Hiruma, Y. *et al.* The structure of the cytochrome p450cam-putidaredoxin complex determined by paramagnetic NMR spectroscopy and crystallography. *J. Mol. Biol.* **425,** 4353–65 (2013).

46. Tripathi, S., Li, H. & Poulos, T. L. Structural basis for effector control and redox partner recognition in cytochrome P450. *Science* **340,** 1227–30 (2013).

47. *Protein Engineering Handbook*. (John Wiley & Sons, 2012).

48. Kazlauskas, R. J. & Bornscheuer, U. T. Finding better protein engineering strategies. *Nat. Chem. Biol.* **5,** 526–9 (2009).

49. Moore, J. C. & Arnold, F. H. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nat. Biotechnol.* **14,** 458–67 (1996).

50. Stemmer, W. P. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci.* **91,** 10747–10751 (1994).

51. Turner, N. J. Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol.* **5,** 567–73 (2009).

52. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485,** 185–94 (2012).

53. Pleiss, J. Systematic Analysis of Large Enzyme Families: Identification of Specificity- and Selectivity-Determining Hotspots. *ChemCatChem* **6,** 944–950 (2014).

54. Hellinga, H. W. Rational protein design: Combining theory and experiment. *Proc. Natl. Acad. Sci.* **94,** 10015–10017 (1997).

55. *Enzyme Catalysis in Organic Synthesis*. (Wiley-VCH Verlag GmbH & Co. KGaA, 2012). doi:10.1002/9783527639861

56. Westphal, R. *et al.* A tailor-made chimeric thiamine diphosphate dependent enzyme for the direct asymmetric synthesis of (S)-benzoins. *Angew. Chem. Int. Ed. Engl.* **53,** 9376–9 (2014).

57. Wijma, H. J. *et al.* Computationally designed libraries for rapid enzyme stabilization. *Protein Eng. Des. Sel.* **27,** 49–58 (2014).

58. Seifert, A. *et al.* Rational design of a minimal and highly enriched CYP102A1 mutant library with improved regio-, stereo- and chemoselectivity. *Chembiochem* **10,** 853–61 (2009).

59. Seifert, A. & Pleiss, J. Identification of selectivity-determining residues in cytochrome P450 monooxygenases: a systematic analysis of the substrate recognition site 5. *Proteins* **74,** 1028–35 (2009).

60. Ba, L., Li, P., Zhang, H., Duan, Y. & Lin, Z. Semi-rational engineering of cytochrome P450sca-2 in a hybrid system for enhanced catalytic activity: insights into the important role of electron transfer. *Biotechnol. Bioeng.* **110,** 2815–25 (2013).

61. Lutz, S. Beyond directed evolution--semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* **21,** 734–43 (2010).

62. Bernhardt, R. & Urlacher, V. B. Cytochromes P450 as promising catalysts for biotechnological application: chances and limitations. *Appl. Microbiol. Biotechnol.* **98,** 6185–203 (2014).

63. Gillam, E. M. J. Engineering cytochrome p450 enzymes. *Chem. Res. Toxicol.* **21,** 220–31 (2008).

64. Ost, T. W. . *et al.* Rational re-design of the substrate binding site of flavocytochrome P450 BM3. *FEBS Lett.* **486,** 173–177 (2000).

65. Lisurek, M., Simgen, B., Antes, I. & Bernhardt, R. Theoretical and experimental evaluation of a CYP106A2 low homology model and production of mutants with changed activity and selectivity of hydroxylation. *Chembiochem* **9,** 1439–49 (2008).

66. Wu, Z.-L., Podust, L. M. & Guengerich, F. P. Expansion of substrate specificity of cytochrome P450 2A6 by random and site-directed mutagenesis. *J. Biol. Chem.* **280,** 41090–100 (2005).

67. Li, Y. *et al.* A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25,** 1051–6 (2007).

68. Eiben, S., Bartelmäs, H. & Urlacher, V. B. Construction of a thermostable cytochrome P450 chimera derived from self-sufficient mesophilic parents. *Appl. Microbiol. Biotechnol.* **75,** 1055–61 (2007).

69. Hlavica, P., Schulze, J. & Lewis, D. F. V. Functional interaction of cytochrome P450 with its redox partners: a critical assessment and update of the topology of predicted contact regions. *J Inorg Biochem* **96,** 279–97 (2003).

70. Munro, A. W., Girvan, H. M. & McLean, K. J. Cytochrome P450--redox partner fusion enzymes. *Biochim. Biophys. Acta* **1770,** 345–59 (2007).

71. Vogel, C., Widmann, M., Pohl, M. & Pleiss, J. A standard numbering scheme for thiamine diphosphate-dependent decarboxylases. *BMC Biochem.* **13,** 24 (2012).

72. Sirim, D., Widmann, M., Wagner, F. & Pleiss, J. Prediction and analysis of the modular structure of cytochrome P450 monooxygenases. *BMC Struct. Biol.* **10,** 34 (2010).

73. Stjernschantz, E. *et al.* Structural rationalization of novel drug metabolizing mutants of cytochrome P450 BM3. *Proteins* **71,** 336–52 (2008).

74. Lewis, D. F. & Hlavica, P. Interactions between redox partners in various cytochrome P450 systems: functional and structural aspects. *Biochim. Biophys. Acta* **1460,** 353–74 (2000).

75. Vondrásek, J., Bendová, L., Klusák, V. & Hobza, P. Unexpectedly strong energy stabilization inside the hydrophobic core of small protein rubredoxin mediated by aromatic residues: correlated ab initio quantum chemical calculations. *J Am Chem Soc* **127,** 2615–9 (2005).

76. Munro, A. W. *et al.* The role of tryptophan 97 of cytochrome P450 BM3 from Bacillus megaterium in catalytic function. Evidence against the 'covalent switching' hypothesis of P-450 electron transfer. *Biochem. J.* **303 (Pt 2,** 423–8 (1994).

77. Gricman, Ł., Vogel, C. & Pleiss, J. Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining. *Proteins* **83**:1593-1603 (2015).

78. Vogel, C., Reusch, W., Pohl, M., Rother, D. & Pleiss, J. BioCatNet: A system for the analysis of sequence-structure-function relationships of protein families. *Manuscr. Submitt.*

79. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **39,** D32–7 (2011).

80. Whitehouse, C. J. C., Bell, S. G. & Wong, L.-L. P450(BM3) (CYP102A1): connecting the dots. *Chem. Soc. Rev.* **41,** 1218–60 (2012).

81. Gricman, Ł. *et al.* Redox partner interaction sites in cytochrome P450 monooxygenases: in silico analysis and experimental validation. *Manuscr. Submitt.* (2015).

82. Shimizu, T., Tateishi, T., Hatano, M. & Fujii-Kuriyama, Y. Probing the role of lysines and arginines in the catalytic function of cytochrome P450d by site-directed mutagenesis. Interaction with NADPH-cytochrome P450 reductase. *J. Biol. Chem.* **266,** 3372–5 (1991).

83. Lin, H., Kenaan, C., Zhang, H. & Hollenberg, P. F. Reaction of human cytochrome P450 3A4 with peroxynitrite: nitrotyrosine formation on the proximal side impairs its interaction with NADPH-cytochrome P450 reductase. *Chem. Res. Toxicol.* **25,** 2642–53 (2012).

84. Auchus, R. J., Worthy, K., Geller, D. H. & Miller, W. L. Probing structural and functional domains of human P450c17. *Endocr. Res.* **26,** 695–703 (2000).

85. Hong, Y. *et al.* Epitope characterization of an aromatase monoclonal antibody suitable for the assessment of intratumoral aromatase activity. *PLoS One* **4,** e8050 (2009).

86. Allorge, D. *et al.* Functional analysis of CYP2D6.31 variant: homology modeling suggests possible disruption of redox partner interaction by Arg440His substitution. *Proteins* **59,** 339–46 (2005).

87. Wen, B. *et al.* Cysteine 98 in CYP3A4 contributes to conformational integrity required for P450 interaction with CYP reductase. *Arch. Biochem. Biophys.* **454,** 42–54 (2006).

88. Bridges, A. Identification of the Binding Site on Cytochrome P450 2B4 for Cytochrome b5 and Cytochrome P450 Reductase. *J. Biol. Chem.* **273,** 17036–17049 (1998).

89. Omata, Y., Dai, R., Smith, S. V, Robinson, R. C. & Friedman, F. K. Synthetic peptide mimics of a predicted topographical interaction surface: the cytochrome P450 2B1 recognition domain for NADPH-cytochrome P450 reductase. *J. Protein Chem.* **19,** 23–32 (2000).

90. Mayuzumi, H. *et al.* Effect of mutations of ionic amino acids of cytochrome P450 1A2 on catalytic activities toward 7-ethoxycoumarin and methanol. *Biochemistry* **32,** 5622–8 (1993).

91. Lee, M.-Y. *et al.* High warfarin sensitivity in carriers of CYP2C9*35 is determined by the impaired interaction with P450 oxidoreductase. *Pharmacogenomics J.* **14,** 343–9 (2014).

92. Minutolo, C. *et al.* Structure-based analysis of five novel disease-causing mutations in 21-hydroxylase-deficient patients. *PLoS One* **6,** e15899 (2011).

93. Kenaan, C., Zhang, H., Shea, E. V & Hollenberg, P. F. Uncovering the role of hydrophobic residues in cytochrome P450-cytochrome P450 reductase interactions. *Biochemistry* **50,** 3957–67 (2011).

94. Russell, R. B. & Barton, G. J. Multiple Protein Sequence Alignment From Tertiary Structure Comparison : Assignment of Global and Residue Confidence Levels. *Proteins* **14,** 309–323 (1992).

95. Haines, D. C. *et al.* A single active-site mutation of P450BM-3 dramatically enhances substrate binding and rate of product formation. *Biochemistry* **50,** 8333–41 (2011).

96. Poulos, T. L. & Howard, A. J. Crystal structures of metyrapone- and phenylimidazole-inhibited complexes of cytochrome P-450cam. *Biochemistry* **26,** 8165–74 (1987).

97. Pikuleva, I. A., Cao, C. & Waterman, M. R. An additional electrostatic interaction between adrenodoxin and P450c27 (CYP27A1) results in tighter binding than between adrenodoxin and p450scc (CYP11A1). *J. Biol. Chem.* **274,** 2045–52 (1999).

98. Stayton, P. S. & Sligar, S. G. The cytochrome P-450cam binding surface as defined by site-directed mutagenesis and electrostatic modeling. *Biochemistry* **29,** 7381–6 (1990).

99. Scheps, D. *et al.* Synthesis of ω-hydroxy dodecanoic acid based on an engineered CYP153A fusion construct. *Microb. Biotechnol.* **6,** 694–707 (2013).

100. Scheps, D., Malca, S. H., Hoffmann, H., Nestl, B. M. & Hauer, B. Regioselective ω-hydroxylation of medium-chain n-alkanes and primary alcohols by CYP153 enzymes from Mycobacterium marinum and Polaromonas sp. strain JS666. *Org. Biomol. Chem.* **9,** 6727–33 (2011).

101. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).

102. Roberts, G. A., Grogan, G., Greter, A., Flitsch, S. L. & Turner, N. J. Identification of a new class of cytochrome P450 from a Rhodococcus sp. *J. Bacteriol.* **184,** 3898–908 (2002).

103. Kurowski, M. A. & Bujnicki, J. M. GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.* **31,** 3305–7 (2003).

104. Ziegler, G. A., Vonrhein, C., Hanukoglu, I. & Schulz, G. E. The structure of adrenodoxin reductase of mitochondrial P450 systems: electron transfer for steroid biosynthesis. *J. Mol. Biol.* **289,** 981–90 (1999).

105. Ziegler, G. A. & Schulz, G. E. Crystal structures of adrenodoxin reductase in complex with NADP+ and NADPH suggesting a mechanism for the electron transfer of an enzyme family. *Biochemistry* **39,** 10986–95 (2000).

106. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **1543,** 408–415 (2000).

107. Lehmann, M. & Wyss, M. Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. *Curr. Opin. Biotechnol.* **12,** 371–375 (2001).

108. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).

109. Forneris, F., Orru, R., Bonivento, D., Chiarelli, L. R. & Mattevi, A. ThermoFAD, a Thermofluor-adapted flavin ad hoc detection system for protein folding and ligand binding. *FEBS J.* **276,** 2833–40 (2009).

110. Robin, A. *et al.* Engineering and improvement of the efficiency of a chimeric [P450cam-RhFRed reductase domain] enzyme. *Chem. Commun. (Camb)*. 2478–80 (2009). doi:10.1039/b901716j

111. Bell, S. G., Harford-Cross, C. F. & Wong, L.-L. Engineering the CYP101 system for in vivo oxidation of unnatural substrates. *Protein Eng. Des. Sel.* **14,** 797–802 (2001).

112. Bell, S. G. *et al.* Engineering cytochrome P450cam into an alkane hydroxylase. *Dalt. Trans.* 2133 (2003). doi:10.1039/b300869j

113. Polizzi, K. M. *et al.* Pooling for improved screening of combinatorial libraries for directed evolution. *Biotechnol. Prog.* **22,** 961–7

114. Paulsen, M. D. & Ornstein, R. L. Predicting the product specificity and coupling of cytochrome P450cam. *J. Comput. Aided. Mol. Des.* **6,** 449–60 (1992).

115. Filipovic, D., Paulsen, M. D., Loida, P. . J., Sligar, S. G. & Ornstein, L. R. Ethylbenzene hydroxylation by cytochrome P450cam. *Biochem. Biophys. Res. Commun.* **189,** 488–495 (1992).

116. Loida, P. J., Sligar, S. G., Paulsen, M. D., Arnold, G. E. & Ornstein, R. L. Stereoselective Hydroxylation of Norcamphor by Cytochrome P450cam: experimental verification of molecular dynamics simultaions. *J. Biol. Chem.* **270,** 5326–5330 (1995).

117. Lehmann, M. *et al.* The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng. Des. Sel.* **15,** 403–411 (2002).

118. Parthasarathy, S. & Murthy, M. R. N. Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng. Des. Sel.* **13,** 9–13 (2000).

119. Daniel, R. M., Danson, M. J. & Eisenthal, R. The temperature optima of enzymes: a new perspective on an old phenomenon. *Trends Biochem. Sci.* **26,** 223–225 (2001).

120. Maves, S. A. & Sligar, S. G. Understanding thermostability in cytochrome P450 by combinatorial mutagenesis. *Protein Sci.* **10,** 161–8 (2001).

121. Salazar, O., Cirino, P. C. & Arnold, F. H. Thermostabilization of a cytochrome p450 peroxygenase. *Chembiochem* **4,** 891–3 (2003).

122. Kumar, S., Sun, L., Liu, H., Muralidhara, B. K. & Halpert, J. R. Engineering mammalian cytochrome P450 2B1 by directed evolution for enhanced catalytic tolerance to temperature and dimethyl sulfoxide. *Protein Eng. Des. Sel.* **19,** 547–54 (2006).

123. Puchkaev, A. V & Ortiz de Montellano, P. R. The Sulfolobus solfataricus electron donor partners of thermophilic CYP119: an unusual non-NAD(P)H-dependent cytochrome P450 system. *Arch. Biochem. Biophys.* **434,** 169–77 (2005).

124. Mandai, T., Fujiwara, S. & Imaoka, S. A novel electron transport system for thermostable CYP175A1 from Thermus thermophilus HB27. *FEBS J.* **276,** 2416–2429 (2009).

125. Weber, E. *et al.* Screening of a minimal enriched P450 BM3 mutant library for hydroxylation of cyclic and acyclic alkanes. *Chem. Commun.* **47,** 944–6 (2011).

126. Peters, M. W., Meinhold, P., Glieder, A. & Arnold, F. H. Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. *J. Am. Chem. Soc.* **125,** 13442–50 (2003).

127. Fasan, R., Chen, M. M., Crook, N. C. & Arnold, F. H. Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting nativelike catalytic properties. *Angew. Chem. Int. Ed. Engl.* **46,** 8414–8 (2007).

128. Meinhold, P., Peters, M. W., Chen, M. M. Y., Takahashi, K. & Arnold, F. H. Direct conversion of ethane to ethanol by engineered cytochrome P450 BM3. *Chembiochem* **6,** 1765–8 (2005).

129. Koshland, D. E. The Key–Lock Theory and the Induced Fit Theory. *Angew. Chemie Int. Ed. English* **33,** 2375–2378 (1995).

130. Chica, R. A., Doucet, N. & Pelletier, J. N. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr. Opin. Biotechnol.* **16,** 378–84 (2005).

131. Mao, W., Rupasinghe, S. G., Zangerl, A. R., Berenbaum, M. R. & Schuler, M. A. Allelic variation in the Depressaria pastinacella CYP6AB3 protein enhances metabolism of plant allelochemicals by altering a proximal surface residue and potential interactions with cytochrome P450 reductase. *J. Biol. Chem.* **282,** 10544–52 (2007).

132. Kabumoto, H., Miyazaki, K. & Arisawa, A. Directed evolution of the actinomycete cytochrome P450moxA (CYP105) for enhanced activity. *Biosci. Biotechnol. Biochem.* **73,** 1922–7 (2009).

133. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. & Fersht, A. R. Strength and co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.* **216,** 1031–44 (1990).

134. Spector, S. *et al.* Rational Modification of Protein Stability by the Mutation of Charged Surface Residues †. *Biochemistry* **39,** 872–879 (2000).

135. Vogt, G., Woell, S. & Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **269,** 631–43 (1997).

136. Gocke, D. *et al.* Rational protein design of ThDP-dependent enzymes-engineering stereoselectivity. *Chembiochem* **9,** 406–12 (2008).

137. Zhao, H. & Arnold, F. H. Combinatorial protein design: strategies for screening protein libraries. *Curr. Opin. Struct. Biol.* **7,** 480–485 (1997).

138. Li, Q. S., Schwaneberg, U., Fischer, P. & Schmid, R. D. Directed evolution of the fatty-acid hydroxylase P450 BM-3 into an indole-hydroxylating catalyst. *Chemistry* **6,** 1531–6 (2000).

139. White, R. E., Miller, J. P., Favreau, L. V & Bhattacharyya, A. Stereochemical dynamics of aliphatic hydroxylation by cytochrome P-450. *J. Am. Chem. Soc.* **108,** 6024–31 (1986).

140. Keizers, P. H. J. *et al.* Metabolic regio- and stereoselectivity of cytochrome P450 2D6 towards 3,4-methylenedioxy-N-alkylamphetamines: in silico predictions and experimental validation. *J. Med. Chem.* **48,** 6117–27 (2005).

141. Chang, Y. T. & Loew, G. H. Molecular dynamics simulations of P450 BM3--examination of substrate-induced conformational change. *J. Biomol. Struct. Dyn.* **16,** 1189–203 (1999).

142. Juhl, P. B., Doderer, K., Hollmann, F., Thum, O. & Pleiss, J. Engineering of Candida antarctica lipase B for hydrolysis of bulky carboxylic acid esters. *J. Biotechnol.* **150,** 474–80 (2010).

143. Wijma, H. J., Marrink, S. J. & Janssen, D. B. Computationally efficient and accurate enantioselectivity modeling by clusters of molecular dynamics simulations. *J. Chem. Inf. Model.* **54,** 2079–92 (2014).

144. Pace, V. *et al.* Structural bases for understanding the stereoselectivity in ketone reductions with ADH from Thermus thermophilus: A quantitative model. *J. Mol. Catal. B Enzym.* **70,** 23–31 (2011).

145. Branco, R. J. F. *et al.* Anchoring effects in a wide binding pocket: the molecular basis of regioselectivity in engineered cytochrome P450 monooxygenase from B. megaterium. *Proteins* **73,** 597–607 (2008).

146. Seifert, A., Antonovici, M., Hauer, B. & Pleiss, J. An efficient route to selective bio-oxidation catalysts: an iterative approach comprising modeling, diversification, and screening, based on CYP102A1. *Chembiochem* **12,** 1346–51 (2011).

147. Benkert, P., Tosatto, S. C. E. & Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins* **71,** 261–77 (2008).

148. Melo, F., Devos, D., Depiereux, E. & Feytmans, E. ANOLEA: a www server to assess protein structures. *ISMB* (1997).

149. Nelson, D. R. Progress in tracing the evolutionary paths of cytochrome P450. *Biochim Biophys Acta* **1814,** 14–8 (2011).

150. Nelson, D. R. *et al.* The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. *DNA Cell Biol* **12,** 1–51 (1993).

151. Sirim, D., Wagner, F., Lisitsa, A. & Pleiss, J. The cytochrome P450 engineering database: Integration of biochemical properties. *BMC Biochem* **10,** 27 (2009).

152. Lewis, D. F. *Cytochromes P450: Structure, function and mechanism*. (1996).

153. Nebert, D. W. & Russell, D. W. Clinical importance of the cytochromes P450. *Lancet* **360,** 1155–62 (2002).

154. Danielson, P. B. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. *Curr. Drug Metab.* **3,** 561–97 (2002).

155. Davydov, D. R., Kariakin, A. A., Petushkova, N. A. & Peterson, J. A. Association of cytochromes P450 with their reductases: opposite sign of the electrostatic interactions in P450BM-3 as compared with the microsomal 2B4 system. *Biochemistry* **39,** 6489–97 (2000).

156. Zawaira, A. *et al.* Exhaustive computational search of ionic-charge clusters that mediate interactions between mammalian cytochrome P450 (CYP) and P450-oxidoreductase (POR) proteins. *Comput Biol Chem* **34,** 42–52 (2010).

157. Bernhardt, R., Kraft, R., Otto, A. & Ruckpaul, K. Electrostatic interactions between cytochrome P-450 LM2 and NADPH-cytochrome P-450 reductase. *Biomed Biochim Acta* **47,** 581–592 (1988).

158. Hasemann, C. A., Kurumbail, R. G., Boddupalli, S. S., Peterson, J. A. & Deisenhofer, J. Structure and function of cytochromes P450: a comparative analysis of three crystal structures. *Structure* **3,** 41–62 (1995).

159. Galleni, M. *et al.* Standard numbering scheme for class B beta-lactamases. *Antimicrob Agents Ch* **45,** 660–663 (2001).

160. Widmann, M., Pleiss, J. & Oelschlaeger, P. Systematic Analysis of Metallo-beta-Lactamases Using an Automated Database. *Antimicrob Agents Ch* **56,** 3481–3491 (2012).

161. Garau, G. *et al.* Update of the standard numbering scheme for class B beta-lactamases. *Antimicrob Agents Ch* **48,** 2347–2349 (2004).

162. Al-Lazikani, B., Lesk, A. M. & Chothia, C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* **273,** 927–948 (1997).

163. Baudry, J., Rupasinghe, S. & Schuler, M. a. Class-dependent sequence alignment strategy improves the structural and functional modeling of P450s. *Protein Eng Des Sel* **19,** 345–53 (2006).

164. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–1 (2010).

165. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res* **33,** D34–8 (2005).

166. Chenna, R. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31,** 3497–3500 (2003).

167. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol.* **7,** e1002195 (2011).

168. Stayton, P. S., Poulos, T. L. & Sligar, S. G. Putidaredoxin Competitively Inhibits Cytochrome b5-Cytochrome P-450cam Association : A Proposed Molecular Model for a Cytochrome P-450cam Electron-Transfer Complex. *Biochemistry* **28,** 8201–8205 (1989).

169. Hardwick, J. P., Songs, B., Huberman, E. & Gonzalezq, F. J. Isolation , Complementary DNA Sequence, and Regulation of Rat Hepatic Lauric Acid o-Hydroxylase (Cytochrome P-450LAo). *J Biol Chem* **262,** 801–810 (1987).

170. Lamb, D. C. & Waterman, M. R. Unusual properties of the cytochrome P450 superfamily. *Philos Trans R Soc L. B Biol Sci* **368,** 20120434 (2013).

171. Sezutsu H, Le Goff G, F. R. Origins of P450 diversity. *Philos Trans R Soc L. B Biol Sci* **368,** 20120428 (2013).

172. Nelson, D. R. A world of cytochrome P450s. *Philos Trans R Soc L. B Biol Sci* **368,** 20120430 (2013).

173. Haines, D. C., Tomchick, D. R., Machius, M. & Peterson, J. A. Pivotal role of water in the mechanism of P450BM-3. *Biochemistry* **40,** 13456–65 (2001).

174. Nagano, S. & Poulos, T. L. Crystallographic study on the dioxygen complex of wild-type and mutant cytochrome P450cam. Implications for the dioxygen activation mechanism. *J. Biol. Chem.* **280,** 31659–63 (2005).

175. Aikensf, J. & Sligar, S. G. Kinetic Solvent Isotope Effects during Oxygen Activation. *J Am Chem Soc* 1143–1144 (1994).

176. Bm, F. P. *et al.* Phenylalanine 393 Exerts Thermodynamic Control over the Heme of Flavocytochrome P450 BM3. *Biochemistry* **74,** 13421–13429 (2001).

177. Ost, T. W. B. *et al.* Oxygen activation and electron transfer in flavocytochrome P450 BM3. *J Am Chem Soc* **125,** 15010–20 (2003).

178. Osváth, S. & Gruebele, M. Proline can have opposite effects on fast and slow protein folding phases. *Biophys J* **85,** 1215–22 (2003).

179. Balbach, J. & Schmid, F. X. in *Mech. protein Fold.* 212–249 (Oxford: Oxford University Press, 2000).

180. Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem* **34,** 167–339 (1981).

181. Schlichting, I. *et al.* The catalytic pathway of cytochrome p450cam at atomic resolution. *Science* **287,** 1615–22 (2000).

182. Haines, D. C., Tomchick, D. R., Machius, M. & Peterson, J. A. Pivotal Role of Water in the Mechanism of P450BM-3†. *Biochemistry* **40,** 13456–13465 (2001).

183. Ost, T. W. *et al.* Phenylalanine 393 exerts thermodynamic control over the heme of flavocytochrome P450 BM3. *Biochemistry* **40,** 13421–9 (2001).

184. Shen, S. J. & Strobel, H. W. Role of Lysine and Arginine Residues of Cytochrome P450 in the Interaction between Cytochrome P4502B1 and NADPH-Cytochrome P450 Reductase. *Arch Biochem Biophys* **304,** 257–265 (1993).

185. Shimizu, T., Tateishi, T., Hatano, M. & Fujii-Kuriyama, Y. Probing the role of lysines and arginines in the catalytic function of cytochrome P450d by site-directed mutagenesis. Interaction with NADPH-cytochrome P450 reductase. *J. Biol. Chem.* **266,** 3372–3375 (1991).

186. Benveniste, I., Lesot, A., Hasenfratz, M.-P., Kochs, G. & Durst, F. Multiple forms of NADPH-cytochrome P450 reductase in higher plants. *Biochem Biophys Res Commun* **177,** 105–112 (1991).

187. Backes, W. L. & Kelley, R. W. Organization of multiple cytochrome P450s with NADPH-cytochrome P450 reductase in membranes. *Pharmacol Ther.* **98,** 221–233 (2003).

188. Mackenzie, F., Cherkesova, T., Grabovec, I. & Strushkevich, N. Structural basis for pregnenolone biosynthesis by the mitochondrial monooxygenase system. *P Natl Acad Sci USA* **108,** 15535–15535 (2011).

189. Burley, S. K. & Petsko, G. A. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science (80-.).* **229,** 23–8 (1985).

190. Unno, M., Shimada, H., Toba, Y., Makino, R. & Ishimura, Y. Role of Arg 112 of Cytochrome P450 cam in the Electron Transfer from Reduced Putidaredoxin. *J Biol Chem* **271,** 17869–17874 (1996).

191. Yano, J. K., Hsu, M.-H., Griffin, K. J., Stout, C. D. & Johnson, E. F. Structures of human microsomal cytochrome P450 2A6 complexed with coumarin and methoxsalen. *Nat Struct Mol Biol* **12,** 822–3 (2005).

192. Ortiz de Montellano, P. R. *Cytochrome P450: Structure, Mechanism, and Biochemistry*. (Kulwer Academic/Plenum Publishers, 2005).

193. Smith, D. A., Ackland, M. J. & Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates Part 1: active site characteristics. *Drug Discov. Today* **2,** 406–414 (1997).

194. Sirim, D., Widmann, M., Wagner, F. & Pleiss, J. Prediction and analysis of the modular structure of cytochrome P450 monooxygenases. *BMC Struct. Biol.* **10,** 34 (2010).

195. Rebholz-Schuhmann, D. *et al.* Automatic extraction of mutations from Medline and cross-validation with OMIM. *Nucleic Acids Res.* **32,** 135–42 (2004).

196. Kuipers, R. *et al.* Novel tools for extraction and validation of disease-related mutations applied to Fabry disease. *Hum. Mutat.* **31,** 1026–32 (2010).

197. Lee, L. C., Horn, F. & Cohen, F. E. Automatic extraction of protein point mutations using a graph bigram association. *PLoS Comput. Biol.* **3,** e16 (2007).

198. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

199. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–42 (2000).

200. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–1 (2010).

201. Sayers, E. A General Introduction to the E-utilities. (2010).

202. Monk, B. C. *et al.* Architecture of a single membrane spanning cytochrome P450 suggests constraints that orient the catalytic domain relative to a bilayer. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 3865–70 (2014).

203. Yu, A., Idle, J. R. & Gonzalez, D. F. J. Polymorphic Cytochrome P450 2D6: Humanized Mouse Model and Endogenous Substrates. (2004).

204. Paine, M. J. I. *et al.* Residues glutamate 216 and aspartate 301 are key determinants of substrate specificity and product regioselectivity in cytochrome P450 2D6. *J. Biol. Chem.* **278,** 4021–7 (2003).

205. Gaffney, D. *et al.* Functional characterisation of the H365Y mutation of the 21-hydroxylase gene in congenital adrenal hyperplasia. *J. Steroid Biochem. Mol. Biol.* **123,** 109–14 (2011).

206. Szklarz, G. D., Ornstein, R. L. & Halpert, J. R. Application of 3-dimensional homology modeling of cytochrome P450 2B1 for interpretation of site-directed mutagenesis results. *J. Biomol. Struct. Dyn.* **12,** 061–78 (1994).

207. Li, Q. S., Ogawa, J., Schmid, R. D. & Shimizu, S. Residue size at position 87 of cytochrome P450 BM-3 determines its stereoselectivity in propylbenzene and 3-chlorostyrene oxidation. *FEBS Lett.* **508,** 249–52 (2001).

208. Liu, J. *et al.* The effect of reciprocal active site mutations in human cytochromes P450 1A1 and 1A2 on alkoxyresorufin metabolism. *Arch. Biochem. Biophys.* **424,** 33–43 (2004).

209. Paulsen, M. D., Filipovic, D., Sligar, S. G. & Ornstein, R. L. Controlling the regiospecificity and coupling of cytochrome P450cam: T185F mutant increases coupling and abolishes 3-hydroxynorcamphor product. *Protein Sci.* **2,** 357–65 (1993).

210. Luo, Z., He, Y. A. & Halpert, J. R. Role of residues 363 and 206 in conversion of cytochrome P450 2B1 from a steroid 16-hydroxylase to a 15 alpha-hydroxylase. *Arch. Biochem. Biophys.* **309,** 52–7 (1994).

211. Delano, W. The PyMOL Molecular Graphics System. (2002).

212. Walsh, A. A., Szklarz, G. D. & Scott, E. E. Human cytochrome P450 1A1 structure and utility in understanding drug and xenobiotic metabolism. *J. Biol. Chem.* **288,** 12932–43 (2013).

213. Sugimoto, H. *et al.* Crystal structure of CYP105A1 (P450SU-1) in complex with 1alpha,25-dihydroxyvitamin D3. *Biochemistry* **47,** 4017–27 (2008).

214. Zhao, B., Lei, L., Sundaramoorthy, M., Kagawa, N. & Waterman, M. R. Crystal Structure of Bovine Steroid of 21-hydroxylase (P450c21). *To be Publ.*

215. Sabbadin, F. *et al.* The 1.5-A structure of XplA-heme, an unusual cytochrome P450 heme domain that catalyzes reductive biotransformation of royal demolition explosive. *J. Biol. Chem.* **284,** 28467–75 (2009).

216. Williams, P. A. *et al.* Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* **305,** 683–6 (2004).

217. Tardy, V. *et al.* Phenotype-genotype correlations of 13 rare CYP21A2 mutations detected in 46 patients affected with 21-hydroxylase deficiency and in one carrier. *J. Clin. Endocrinol. Metab.* **95,** 1288–300 (2010).

218. Bui, S. H. *et al.* Unusual spectroscopic and ligand binding properties of the cytochrome P450-flavodoxin fusion enzyme XplA. *J. Biol. Chem.* **287,** 19699–714 (2012).

219. Domanski, T. L., Liu, J., Harlow, G. R. & Halpert, J. R. Analysis of four residues within substrate recognition site 4 of human cytochrome P450 3A4: role in steroid hydroxylase activity and alpha-naphthoflavone stimulation. *Arch. Biochem. Biophys.* **350,** 223–32 (1998).

220. Sawada, Y. & Ayabe, S. Multiple mutagenesis of P450 isoflavonoid synthase reveals a key active-site residue. *Biochem. Biophys. Res. Commun.* **330,** 907–13 (2005).

221. Zhao, B. *et al.* Different binding modes of two flaviolin substrate molecules in cytochrome P450 158A1 (CYP158A1) compared to CYP158A2. *Biochemistry* **46,** 8725–33 (2007).

222. Kahn, R. A., Le Bouquin, R., Pinot, F., Benveniste, I. & Durst, F. A conservative amino acid substitution alters the regiospecificity of CYP94A2, a fatty acid hydroxylase from the plant Vicia sativa. *Arch. Biochem. Biophys.* **391,** 180–7 (2001).

223. Xue, L., Zgoda, V. G., Arison, B. & Almira Correia, M. Structure–function relationships of rat liver CYP3A9 to its human liver orthologs: site-directed active site mutagenesis to a progesterone dihydroxylase. *Arch. Biochem. Biophys.* **409,** 113–126 (2003).

224. Melet, A. *et al.* Analysis of human cytochrome P450 2C8 substrate specificity using a substrate pharmacophore and site-directed mutants. *Biochemistry* **43,** 15379–92 (2004).

225. Attia, T. Z. *et al.* Effect of cytochrome P450 2C19 and 2C9 amino acid residues 72 and 241 on metabolism of tricyclic antidepressant drugs. *Chem. Pharm. Bull. (Tokyo).* **62,** 176–81 (2014).

226. Mast, N., Zheng, W., Stout, C. D. & Pikuleva, I. A. Binding of a cyano- and fluoro-containing drug bicalutamide to cytochrome P450 46A1: unusual features and spectral response. *J. Biol. Chem.* **288,** 4613–24 (2013).

227. Wang, L. H. *et al.* Identification of thromboxane A2 synthase active site residues by molecular modeling-guided site-directed mutagenesis. *J. Biol. Chem.* **271,** 19970–5 (1996).

228. Toporkova, Y. Y. *et al.* Structure-function relationship in the CYP74 family: conversion of divinyl ether synthases into allene oxide synthases by site-directed mutagenesis. *FEBS Lett.* **587,** 2552–8 (2013).

229. Lewis, D. F. V. & Hlavica, P. Interactions between redox partners in various cytochrome P450 systems: functional and structural aspects. *Biochim. Biophys. Acta - Bioenerg.* **1460,** 353–374 (2000).

230. Zhao, C. *et al.* Cross-linking mass spectrometry and mutagenesis confirm the functional importance of surface interactions between CYP3A4 and holo/apo cytochrome b(5). *Biochemistry* **51,** 9488–500 (2012).

231. Koo, L. S., Immoos, C. E., Cohen, M. S., Farmer, P. J. & Ortiz de Montellano, P. R. Enhanced Electron Transfer and Lauric Acid Hydroxylation by Site-Directed Mutagenesis of CYP119. *J. Am. Chem. Soc.* **124,** 5684–5691 (2002).

232. Fleuren, W. W. M. & Alkema, W. Application of text mining in the biomedical domain. *Methods* **74,** 97–106 (2015).

233. Larsen, P. O. & von Ins, M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* **84,** 575–603 (2010).

234. Björk, B.-C. *et al.* Open access to the scientific journal literature: situation 2009. *PLoS One* **5,** e11273 (2010).

235. Jensen, L. J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* **7,** 119–29 (2006).

236. Sevrioukova, I. F., Li, H., Zhang, H., Peterson, J. a & Poulos, T. L. Structure of a cytochrome P450-redox partner electron-transfer complex. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 1863–8 (1999).

237. Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22,** 195–201 (2006).

238. Delano, W. The PyMOL Molecular Graphics System. *available at:http://www.pymol.org* (2002).

239. Rontani, J.-F. & Aubert, C. Trimethylsilyl transfer during electron ionization mass spectral fragmentation of some omega-hydroxycarboxylic and omega-dicarboxylic acid trimethylsilyl derivatives and the effect of chain length. *Rapid Commun. Mass Spectrom.* **18,** 1889–95 (2004).

240. Yasutake, Y., Nishioka, T., Imoto, N. & Tamura, T. A single mutation at the ferredoxin binding site of P450 Vdh enables efficient biocatalytic production of 25-hydroxyvitamin D(3). *Chembiochem* **14,** 2284–91 (2013).

241. Van Vugt-Lussenburg, B. M. A. *et al.* Identification of critical residues in novel drug metabolizing mutants of cytochrome P450 BM3 using random mutagenesis. *J. Med. Chem.* **50,** 455–61 (2007).

242. Sevrioukova, I. F., Hazzard, J. T., Tollin, G. & Poulos, T. L. The FMN to heme electron transfer in cytochrome P450BM-3. Effect of chemical modification of cysteines engineered at the FMN-heme domain interaction site. *J. Biol. Chem.* **274,** 36097–106 (1999).

243. Tsotsou, G. E., Sideri, A., Goyal, A., Di Nardo, G. & Gilardi, G. Identification of mutant Asp251Gly/Gln307His of cytochrome P450 BM3 for the generation of metabolites of diclofenac, ibuprofen and tolbutamide. *Chemistry* **18,** 3582–8 (2012).

244. Urlacher, V. B., Makhsumkhanov, A. & Schmid, R. D. Biotransformation of beta-ionone by engineered cytochrome P450 BM-3. *Appl. Microbiol. Biotechnol.* **70,** 53–9 (2006).

245. Whitehouse, C. J. C. *et al.* Structural basis for the properties of two single-site proline mutants of CYP102A1 (P450BM3). *Chembiochem* **11,** 2549–56 (2010).

246. Chen, Z., Ost, T. W. B. & Schelvis, J. P. M. Phe393 mutants of cytochrome P450 BM3 with modified heme redox potentials have altered heme vinyl and propionate conformations. *Biochemistry* **43,** 1798–808 (2004).

247. Lo, K. K.-W., Wong, L.-L. & Hill, H. A. O. Surface-modified mutants of cytochrome P450cam: enzymatic properties and electrochemistry. *FEBS Lett.* **451,** 342–346 (1999).

248. Zou, M. *et al.* Mutation prediction by PolyPhen or functional assay, a detailed comparison of CYP27B1 missense mutations. *Endocrine* **40,** 14–20 (2011).

249. Li, D., Dammer, E. B. & Sewer, M. B. Resveratrol stimulates cortisol biosynthesis by activating SIRT-dependent deacetylation of P450scc. *Endocrinology* **153,** 3258–68 (2012).

250. Duan, Y. *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **35,** 1999–2012 (2003).

251. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4,** 435–447 (2008).

252. Seifert, A., Tatzel, S., Schmid, R. D. & Pleiss, J. Multiple molecular dynamics simulations of human p450 monooxygenase CYP2C9: the molecular basis of substrate binding and regioselectivity toward warfarin. *Proteins* **64,** 147–55 (2006).

253. Wang, Y. *et al.* PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **37,** W623–33 (2009).

254. Krieger, Elmar, G. V. & Spronk, C. YASARA–Yet Another Scientific Artificial Reality Application. *YASARA.org* (2013).

255. Dupradeau, F.-Y. *et al.* The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Phys. Chem. Chem. Phys.* **12,** 7821–39 (2010).

256. Vanquelef, E. *et al.* R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **39,** W511–7 (2011).

257. Bayly, C. I., Cieplak, P., Cornell, W. D. & Kollman, P. A. ARTICLES A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving. *J. Phys. Chem.* **97,** 10269–10280 (1993).

258. Case, D. A. *et al.* AMBER 10. *Univ. California, San Fr.* (2008).

259. Parrinello, M. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182 (1981).

260. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **52,** 255–268 (2006).

261. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089 (1993).

262. Chronopoulou, E. G. & Labrou, N. E. Site-saturation mutagenesis: a powerful tool for structure-based design of combinatorial mutation libraries. *Curr. Protoc. Protein Sci.* **Chapter 26,** Unit 26.6 (2011).

Declaration

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Stuttgart, May 28 2015

_____

Łukasz Gricman