Protein and DNA technologies for functional expression of membrane-associated cytochromes P450 in bacterial cell factories

Vazquez Albacete, Dario

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Darío Vázquez-Albacete

Protein and DNA technologies for functional expression of membrane-associated cytochromes P450

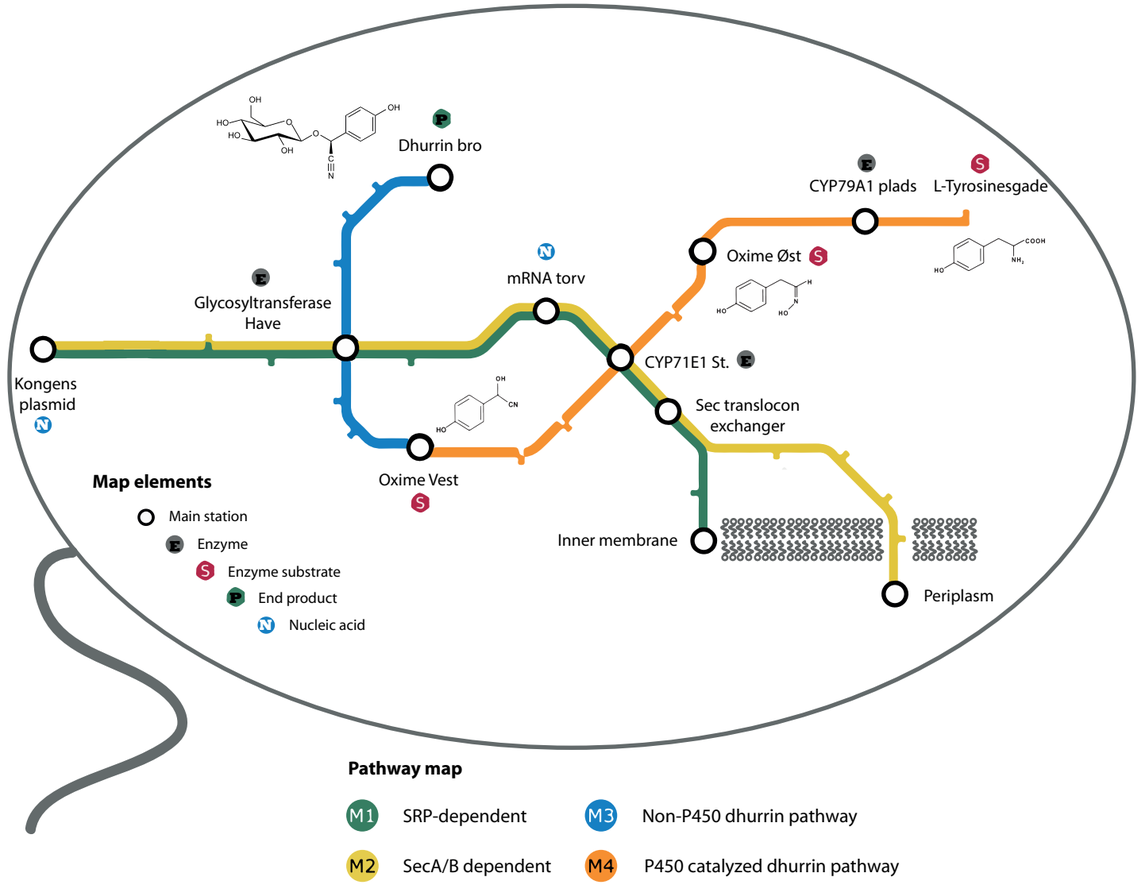
Novo Nordisk Foundation
Center for Biosustainability
Technical University of Denmark

www.biosustain.dtu.dk

DTU

Protein and DNA technologies for functional expression of membrane-associated cytochromes P450 in bacterial cell factories

Novo Nordisk Foundation Center for Biosustainability



Darío Vázquez-Albacete
PhD thesis
August 2016

DTU

DTU Biosustain
The Novo Nordisk Foundation Center for Biosustainability

Preface

This thesis is written as a partial fulfilment of the requirements to obtain a Ph.D. degree at the Technical University of Denmark. This thesis was carried out at the Novo Nordisk Foundation Center For Biosustainibility, Department of DTU Biosustain, Technical University of Denmark from March 2013 to February 2016 under the supervision of Morten Nørholm. This Ph.D. project was funded by The Novo Nordisk Foundation and a Ph.D. grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme [FP7-People-2012-ITN], under grant agreement no. 317058, "BACTORY".

Dario Vazquez-Albacete

Hørsholm, June 2016

Fortitudine vincimus (By endurance we conquer)

Sir Ernest Shackleton family motto

Abstract

The heavy dependence and massive consumption of fossil fuels by humans is changing our environment very rapidly. Some of the side effects of industrial activity include the pollution of the natural resources we rely on, and the reduction of biodiversity. Some chemicals found in nature exhibit great potential as medicines, fuels or food for humans. Plants conquered different environments thereby developing adaptation strategies based on the biosynthesis of a myriad of compounds. Unfortunately they are present in small amounts in plants and are too complex and to produce by organic chemical synthesis. In most of biosynthetic pathways leading to these chemicals the cytochrome P450 enzyme family (P450s) is responsible for their final functionalization. However, the membrane-bound nature of P450s, makes their expression in microbial hosts a challenge. In order to meet the global demand for these natural compounds without compromising sustainability, biological production needs to substitute the traditional manufacturing methods. Thus, new methodologies for expression and characterization of P450 enzymes are in great need.

This thesis explores state-of-the-art techniques at the core of membrane protein, metabolic engineering and protein engineering to provide new solutions to the P450 expression bottleneck in bacteria. The work primarily focuses on developing a fluorescence high-throughput platform to easily assess proper folding and expression levels of plant cytochromes P450. The platform has been designed to fit in metabolic engineering and structural biology applications. Furthermore in this thesis a systematic engineering rationale is proposed to improve P450 expression. For this, a new set of N-terminal tags has been developed in order to provide a streamlined optimization scheme for P450 expression. The application of these N-terminal tags has been also tested to elucidate the structure of the plant cytochrome P450 CYP79A1.

The present work demonstrates the usefulness of the abovementioned technologies to optimize P450 expression for biotechnological applications. The thesis provides new P450 engineering guidelines and serves as platform to improve performance of microbial cells, thereby boosting recombinant production of complex plant P450-derived biochemicals. The knowledge generated, could guide future reconstruction of functional plant metabolic pathways leading to high valuable chemicals. This work is in the foundations of sustainability, as it contributes to find alternatives that limits or relief exploitation of scarce natural resources vital for the survival of our future generations.

Dansk resumé

Menneskets afhængighed af fossile brændstoffer er hurtigt på vej til at ændre vores planet. Bivirkningerne af den industrielle revolution inkluderer forurening og en reduceret biodiversitet. Planter erobrer og tilpasser sig forskellige miljøer vha. en lang række forskellige stoffer og nogle af de kemikalier vi finder i naturen har stort potentiale indenfor medicin, brændstoffer og fødevarer. Desværre er disse stoffer svært tilgængelige: de er tilstede i meget små mængder i planter og er næsten umulige at syntetisere kemisk. De fleste biosynteseveje, der fører til disse stoffer, involverer cytokrom P450 enzymer. Desværre er disse enzymer svære at udtrykke i mikroorganismer. For at møde et fremtidigt behov for produktion af naturstoffer uden at kompromittere bæredygtighed, må biologiske produktionsmetoder erstatte de traditionelle fabrikationsmetoder. Derfor er der et stort behov for nye metoder til udnyttelsen af fx sådanne P450 enzymer.

Denne afhandling udforsker brugen af moderne teknikker indenfor membranproteiner, metabolic og protein engineering og beskriver nye løsninger på problemerne med udnyttelsen af P450 enzymer. Arbejdet har primært været fokuseret på at udvikle en platform for produktion af korrekt foldede P450 enzymer til brug indenfor strukturbiologi og metabolic engineering. Til dette formål har vi bla. udviklet brugen af små peptider der forbedrer udtrykkelsen af P450ere og dermed deres udnyttelse indenfor bioteknologiske applikationer. Afhandlingen beskriver guidelines og en platform til at forbedre mikrobielle cellefabrikker, der producerer P450-afledte biokemikalier og den genererede viden vil forbedre vores muligheder for at producere planteafledte højværdistoffer. Arbejdet danner dermed grundlaget for bæredygtige alternativer, der kan sikre overlevelsen af fremtidige generationer.

Acknowledgements

I would like to acknowledge first of all my family, who gave me the opportunity and resources to develop myself as a scientist. My mother Lourdes Albacete who supported me unconditionally, educated me to develop my creativity, my emotional intelligence, respect for others, and gave me freedom of choice in all aspects of my life. All these values have shaped my social character, boosted my collaborative and creative mind-set. This thesis represents a milestone in our lives that I'm pleased to specially dedicate it to her. To my father Manuel Vazquez whom awakened my curiosity for science since I was a child. But most importantly he taught me three skills that I found indispensable to follow my own dreams 1) Perseverance; through the motto by endurance we conquer, namely, to not give up on your goals regardless of the adversities we find the way 2) Disciple; follow strict methods and codes to pursue the goals that we want to meet 3) Anticipate possible setbacks; this way you can be prepared for the worst scenario in advance, and always have a B plan. I would also like to acknowledge my aunt Carmen Albacete for supporting me in all bad moments of my student life and professional career. Thanks to her I could focus to finish my bachelor degree. To my sister Arantxa Vazquez, for always being by my side and encouraging me to move on. Special thanks to the rest of my family.

I want to dedicate special thanks to Denise Oró-Bozzini, my fiancée, for many reasons; First of all for supporting me throughout my whole career. Second, for her feedback, not only on science-related matters but also at personal level in the moments I needed it. And third, for her love and perseverance that has kept our relationship untouched despite being more than 2000 kilometres away from each other for nearly two years and a half. She is and always will be my science and personal hero. Also acknowledge her family for all their good wishes and positive thrill during all these years.

I have to thank all people involved in the BacTory program, in particular my supervisor Morten Nørholm for believing in me, for giving me a unique opportunity to develop myself as a scientist, for the enriching science discussions we had during the whole project, for his availability, proximity and outstanding creativity, which has been of great inspiration for me. Thanks to Søren Molin who provided great guidance during the PhD, for giving us the opportunity to meet great scientists, and for believing in the Danish-Spanish alliance. Also thank my colleagues of the BacTory program, especially my office mates Sofie, Mafalda, Mikkel, Isotta and Xiao Chen. But also to Patricia for being one my closest friends during my entire thesis, for having so much fun, and noise, together in

the office. It has been a great pleasure. I also have to acknowledge the positive thrill of Morten's group members and their good team work skills. In particular, Virginia Martinez who always gave me good feedback of my project, helped me to develop critical thinking on and be one of the best colleagues in my group.

Finally I would like to thank all members of the Society of Spanish Researchers in Denmark (*Spanske Forskere I Danmark*) for facilitating the settlement of Spanish scientists in Denmark and for their contribution to the development of my collaborative skills, which were indispensable during my PhD. I especially want to thank the most active members of this association for believing in a project in which two different nations can work together to build a better future for our societies.

Abbreviations

E. coli (*Escherichia coli*)

DNA (Deoxyribonucleic acid)

DBTA (Design-Build-Test-Analyze)

PCR (Polymerase chain reaction)

TM (Transmembrane segment)

IMP (Inner membrane proteins)

SRP (Signal Recognition Particle)

GFP (Green Fluorescent Protein)

IPTG (Isopropyl β-D-1-thiogalactopyranoside)

NADPH (Nicotinamide adenine dinucleotide phosphate)

P450 (cytochrome P450)

CPR (cytochrome P450 reductase)

SDS-PAGE (SDS Polyacrylamide gel electrophoresis)

SRS (Substrate recognition site)

SCR (Structural conserved region)

ATP (Adenosine Triphosphate)

GTP (Guanosine Triphosphate)

Contents

1. Biosustainability: Applications of biotechnology for sustainable development

In general terms, sustainability can be defined as the endurance of systems and processes over time. Sustainability is understood as a system consisting of three differentiated elements; ecology and environment, economy, and society (Fig 1). An important core of sustainability is though ensuring resources of future generations, englobed in the social circle and implies a long-term thinking in sustainable development initiatives. For this reason sustainable development requires research in science and technology to find solutions that can satisfy environmental, economic and social challenges.

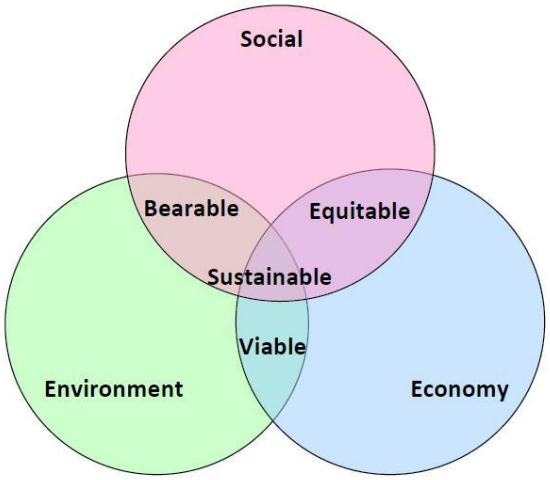


Fig 1. The three dimensions of sustainability

Biosustainability comes into scene as a more specific term in which biotechnology represents a major socioeconomic driver of sustainable development. Biotechnology in its broad definition is the use of biological systems to develop processes, products and technological applications for the benefit of society. Here it is discussed how biotechnology has brought important advances that fulfil the three key elements of sustainability. Early examples of biotechnology can be seen in brewing or baking, in which a microbe is responsible for converting relatively cheap materials into added value products. However, it wasn't until the 20th century when two examples of biotechnological applications gave rise to an entire new industry.

The first example is not the discovery, but the development of Penicillin as a commercial drug (Demain 2006). Initially discovered in 1929 by Alexander Flemming, Penicillin was not possible to produce in bulk until 1944 due to the low productivity of the first wild type mold. The need for massive number of doses of penicillin boosted the development of the fermentation technology and strain engineering techniques such as mutagenesis (Demain 2006). The optimization of penicillin production was led by Florey, Chain and colleagues whom were not initially acknowledged by their achievements, which saved millions of lives

(Table 1). Nevertheless this case settled down the basis of the production of high valuable compounds, and more particularly secondary and primary metabolites using microbes as tiny factories (cell factories).

Table 1. Production of penicillins

Year	Production (kg)	Cost ($/kg)
1945	2,300	11,000
1963	3,000,000	150
1978	15,000,000	18.50
1992	22,000,000	_
1995	31,000,000	4.5

A second illustrative example relies on the insulin production technology. Although this hormone was discovered in 1922 by the Canadians Banting and Best, major breakthroughs in the use of insulin were made in Denmark. First the pioneer couple August and Marie Krogh whom founded the Nordisk Insulinlaboratorium (later Novo Nordisk) imported the rights to commercialize and develop insulin in Denmark. The hormone was extracted from porcine pancreas for decades, however, patients developed antibodies against xenogenic insulin. This motivated extensive research in insulin formulation, and purity to decrease the side effects with the contribution of several Danish characters. On the other hand Revolutionary discoveries in recombinant DNA technology in the 1970's led to the establishment of the major production hosts (*Escherichia coli*, *Bacillus subtilis* or *Saccharomyces cerevisiae*), which in combination with the previous knowledge gave rise in 1987 to the industrial production of the first humanized insulin in yeast cells (Demain 2006).

These two pragmatic examples kicked-off the biotechnology era, and in particularly the use of microbes as cell factories for contributing to the economic, social and environmental development. In the present times, growing concerns over the steady consumption of fossil fuels threatens economic stability and the environment. This is motivating the development of sustainable processes for the production multiple materials from renewable resources. Not only oil dependency, but also the complexity of certain chemicals found in natural ecosystems, motivates their sustainable exploitation. For these reason the application of cell factories for sustainable development entered the spot light for companies and political agendas.

2. Cell Factories and metabolic engineering

A whole microbial, mammalian or plant cell with the ability to produce a certain chemical is considered as a cell factory wherein internal metabolic pathways give the properties to convert one or several substrates into the product of interest instead of the traditional single reaction concept. Microorganisms are particularly versatile cell factories due to their scalability, fast growth rate, well-known physiology and easy manipulation. Examples of compounds produced are antibiotics, biofuels, fine chemicals and drugs (Lee, Na et al. 2012, Murphy 2012).

For optimizing cell factories it is crucial to understand metabolic networks from a systems perspective and develop tools to manipulate these systems on our will. However, the lack of understanding on complex metabolic, gene regulatory and signalling networks leaves us far from this achievement. Metabolic engineering represents a technological framework to accelerate and modify existing pathways for the optimal production of desired products using cell factories efficiently (Nielsen, Fussenegger et al. 2014). More specifically are a handful of techniques to engineer cell factories for the biological manufacturing of chemicals and biopharmaceuticals (Stephanopoulos 2012). On the other hand the means by which we learn about biological parts and systems to ultimately engineer metabolic pathways is known as synthetic biology. The goal of synthetic biology is to simplify biological engineering through the application of engineering principles and design.

2.1. Cell factory development cycle: Design-build-test-analyze (DBTA)

In a traditional chemical factory it is possible to reconfigure unit operations using design, process control systems software and predict the outcome by mathematical modelling, which help maximizing production, profits and safety. Unfortunately the complexity of biological systems makes it difficult to accurately control metabolic pathways of cell factories to achieve a certain production goal or profitability. In this context the implementation of the DBTA cycle, inspired by electronic engineering principles, comes as useful approach for cell factory optimization (Fig. 2). Often the DNA parts constituting the genetic elements encoding for a metabolic pathway, can be modified through several DBTA cycles upon rendering an optimal production yield for example. This can be achieved through different DNA editing and assembly methods not discussed in this thesis (Casini, Storch et al. 2015, Cavaleiro, Kim et al. 2015).

The most prominent example of the DBTA cycle implementation can be found in the production of the semi-synthetic Artemisinin, a blockbuster anti-malaria drug with 1.5 billion dollar annual sales worldwide (Paddon and Keasling 2014). First, the isoprene metabolic pathway leading to the artemisinin terpene backbone precursor, amorphadiene, was expressed in *E. coli*. However this strategy turned out to have limited impact on amorphadiene production. A new DBTA cycle iteration was carried out for

introducing the same pathway from yeast, increasing the titers substantially. Finally the enzyme responsible for converting isoprenoids in amorphadiene, *Artemisina annua* amorphadiene synthase (ADS), was optimized for codon usage in *E. coli*, consequently increasing the yields. Several iteration cycles followed with different enzymes of the heterologous pathways involved in amorphadiene biosynthesis leading to final titers of 25 g amorphadiene /L in *E. coli* compared to 0.5g/L of the previous optimization steps and proving the fast development cycle of cell factories.

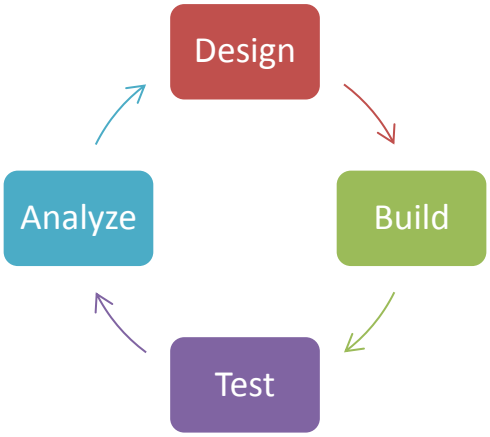


Fig. 1 The Design-Build-Test-Analyze (DBTA) cycle can be used to accelerate the development of microbial cell factories.

2.2. *Escherichia coli* as cell factory

Escherichia coli is a common inhabitant found in the mammalian intestines and is one of the best studied organisms. While the origin of *E. coli* K-12 strain, named by Delbrück and Luria in 1942, is clear, the early history the *E. coli* strain B is less well established. *E. coli* has been a workhorse in fundamental studies and biotechnological applications. Recent advances in systems biology have been changing the way biological studies are performed. The strategies, methods, and tools developed in *E. coli* permitted a deep understanding of it not only as a model system but also for the development of cell factories for production of relevant compounds (Chang, Eachus et al. 2007). Laboratory strains of *E. coli* derive mostly from non-pathogenic K-12 or B strains. In the so-called *KEIO collection*, each non-essential gene of the K-12 strain BW25113 was deleted allowing for a deeper understanding of *E. coli* as a system (Baba, Ara et al. 2006). Strain K-12 has been mainly subjected to metabolic studies whereas B strains received initially less attention and later became the gold standard for protein overexpression (Baumler, Peplinski et al. 2011). However, the genome of two *E. coli* strains of the B lineage, REL606 and BL21(DE3) have been recently analyzed by several "*omic*" techniques including sequencing and compared to the K-12 strain shedding light on essential differences between both strains (Yoon, Han et

al. 2012). All knowledge generated during decades has allowed for the establishment of *E. coli* as one of the most versatile and robust hosts or *chassis* for the development of microbial cell factories.

2.3. *Escherichia coli* expression strains

One of the common problems found in high density cultures of *E. coli* with glucose as a carbon source, is acetate accumulation. Acetate has a major inhibitory effect on cell growth and production of heterologous proteins (Eiteman and Altman 2006). B strains tend to accumulate less acetate than K-12 strains during high cell density cultivation making them more appropriate for protein production (Eiteman and Altman 2006). This is why B strains have been traditionally used for overproduction of heterologous and endogenous proteins because they display faster cell growth and lower production of acetate than K-12 strains.. Analysis of genome sequences of REL606 and BL21(DE3) does not reveal difference in genes involved in the central carbon or acetic acid metabolism (Yoon, Han et al. 2012). Thus, it is possible that the metabolic genes are under different regulation in these two *E. coli* strains. The most popular expression strain has become BL21(DE3), which harbors a genome-integrated T7 RNA polymerase under the control of the *lacUV5* promoter. The T7 RNA polymerase is found in the T7 phage as it is responsible for self-replication of the phage DNA. The advantages of the T7 RNA polymerase are the high selectivity for its promoters, which do not occur naturally in *E. coli*, and also the remarkably fast transcription rate compared to other polymerases (Studier and Moffatt 1986). In addition, B derived strains are defective in *lon* and *ompT* proteases, a feature that can reduce degradation of heterologous proteins (Terpe 2006). Several commercial strains have been developed for gene expression from BL21(DE3) and K-12 (**Table 2**). Important examples are *Rosetta*, that carries a plasmid that supplies tRNAs complementary to rare codons in order to enhance expression of eukaryotic genes or the *Origami* strain that facilitates disulfide bond formation (Terpe 2006).

Table 2. Some *E. coli* strains most frequently used for heterologous protein production and their key features (Terpe 2006)

***E. coli* strain**	**Derivative**	**Key features**
BL21	B834	Deficient in *lon* and *ompT* proteases
JM 83	K-12	Usable for secretion of recombinant proteins into the periplasm
Origami B	BL21	*trxB/gor* mutant; greatly facilitates cytoplasmic disulfide bond formation
Rosetta	BL21	Enhances the expression of eukaryotic proteins that contain codons rarely used in E. coli: AUA, AGG, AGA, CGG, CUA, CCC, and GGA; deficient in *lon* and *ompT* proteases

Often over-production of certain proteins is lethal or toxic to host cells. Furthermore, leaky expression of T7 polymerase from the *lacUV5* promoter is problematic for expression of toxic genes (Miroux and Walker 1996). This is the case for many hydrophobic membrane proteins, which represent around 50% of all drug targets and are of major pharmaceutical and biotechnological interest. Moreover high levels of membrane protein are needed for structural studies, making membrane expression in *E. coli* a main bottleneck for further studies. Because saturation of the translation machinery may occur due to high expression levels, several systems have been designed to tune heterologous gene expression and to limit basal expression of the T7 RNA polymerase (Studier and Moffatt 1986). For example the pLysS and pLysE plasmids expresses different levels of the enzyme T7 lysozyme – a natural inhibitor of the T7 RNA Polymerase (Zhang and Studier 1997). These plasmids can be introduced in BL21(DE3) and are commercially available as BL21(DE3)pLysS and BL21(DE3)pLysE.

2.3.1. *The Walker strains C41(DE3) and C43(DE3)*

In 1996 Miroux and Walker isolated *E. coli* B strains that had evolved to tolerate production of toxic proteins. These strains called C41(DE3) and C43(DE3) are also known as the Walker strains, displayed reduced toxicity toward several membrane proteins (Miroux and Walker 1996). Recent analysis at genome level and experimental validation in C41(DE3) and C43(DE3) revealed that the key of toxicity tolerance relies on); **1)** diminished activity of the *lacUV5* promoter due to mutations, and **2)** mutations in the *lacI* repressor that makes them less sensitive to the inducer (Fig. 3) (Kwon, Kim et al. 2015). This reduces T7 RNA polymerase levels, consequently a more gradual expression of the target gene is achieved (Kwon, Kim et al. 2015).

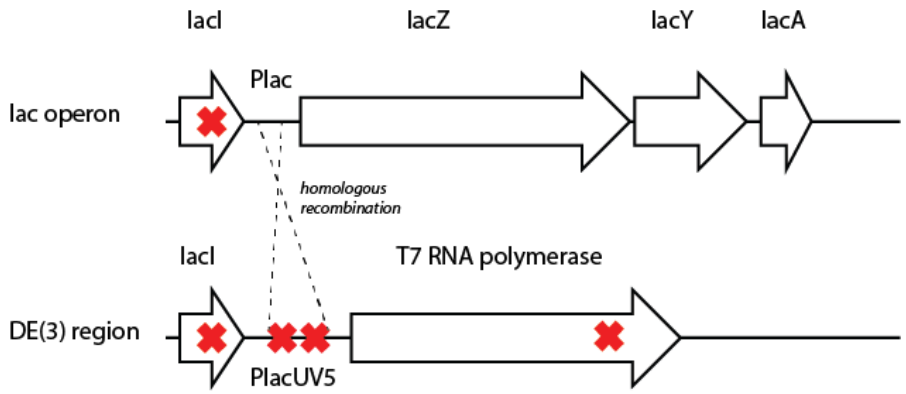


Fig. 3 Hotspot map of mutations identified in the C41(DE3) and C43(DE3) that make *E. coli* cells tolerant to the overexpression of membrane proteins. Red crosses indicate mutated spots. Adapted from (Kwon, Kim et al. 2015).

2.3.2. *Strains with titratable promoters*

Schlegel and colleagues observed that lower RNA polymerase levels could lead to high amount of properly folded proteins, particularly the ones encoding for membrane and periplasmic proteins (Schlegel, Rujas et al. 2013). The concept that less T7 RNA polymerase can actually lead to higher total amounts of properly folded protein motivated the development of the Lemo21(DE3) strain (Schlegel, Rujas et al. 2013). The working foundations of the Lemo21(DE3) rely on a plasmid-based system called *pLemo* in which the T7 lysozyme expression is under the control of the titratable rhamnose promoter (P*rhaBAD*). Upon addition of rhamnose in cell culture media, T7 lysozyme is produced providing a negative regulation of the T7 RNA polymerase in the BL21(DE3) strain (Wagner, Klepsch et al. 2008). Rhamnose is then gradually catabolized by *E. coli*, gradually limiting RNA polymerase inhibition, thus allowing for progressive expression of the target gene under the control of the T7 promoter in an expression vector (Fig. 4). This also reduces the amount of protein aggregates as observed by a decrease of the IpbA chaperones (Schlegel, Rujas et al. 2013). This system has paved the way for structural elucidation of highly relevant drug targets, for example the sodium/proton antiport channel (Lee, Kang et al. 2013).

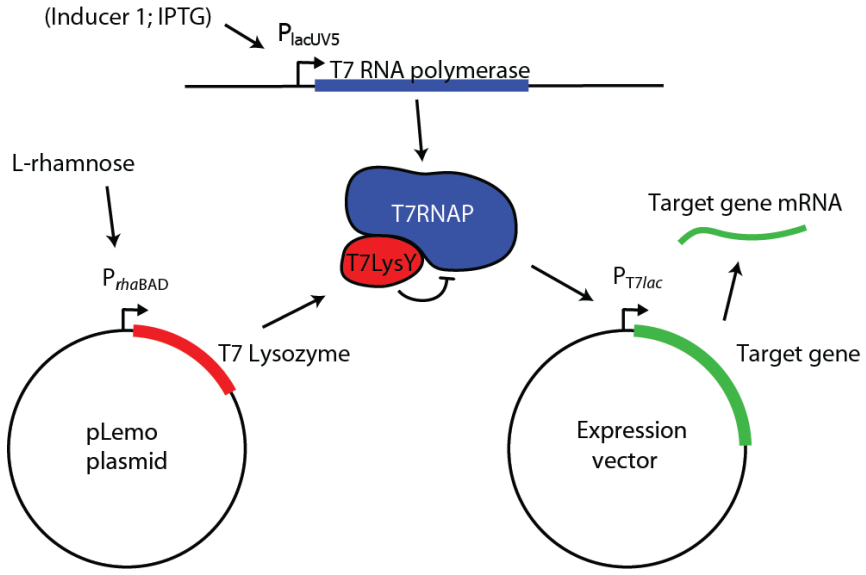


Fig. 4. Lemo21(DE3) working scheme. T7 lysozyme under the control of the P*rhaBAD* represses the T7 RNA polymerase. When rhamnose is gradually catabolized T7 RNA polymerase transcribes the target gene upon addition of IPTG inducer. Adapted from (Schlegel, Rujas et al. 2013).

The *rhaBAD* operon was initially described in 1993 by Egan and Shlief (Egan and Schleif 1993). The promoter is regulated by two activators, RhaS and RhaR belonging to the same transcription unit oriented in the opposite direction of *rhaBAD*. When rhamnose is available, RhaR binds to the *rhaP RS* promoter and

activates the production of additional RhaR and RhaS activators. RhaS and rhamnose in turn bind to *rhaBAD* which is then activated (Fig. 5).

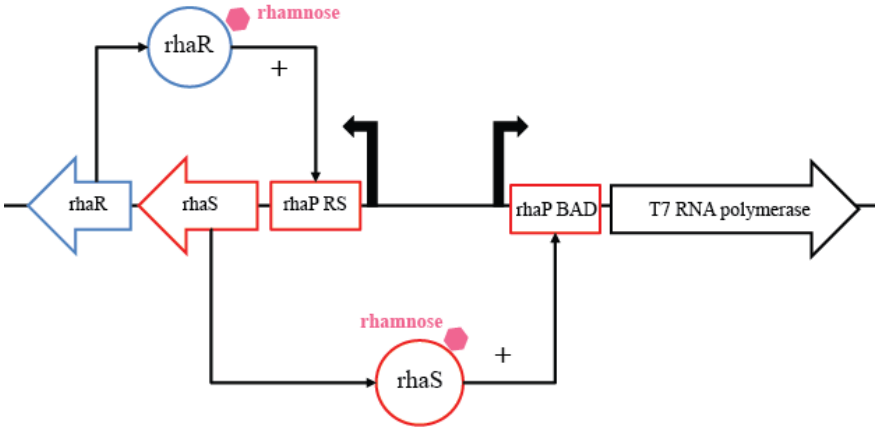


Fig. 5. Schematic representation of the rhamnose operon. The tandem composed by the rhamnose promoter (rha ABD) expression the T7 RNA polymerase is integrated in the genome in the KRX strain.

This promoter has been subjected to several optimization rounds in order to make it stronger and more titratable (Wegerer, Sun et al. 2008). Commercial strains using the rhamnose promoter in other ways has been used for expression of toxic genes. For example the K-12 derived strain KRX, which harbors a genome-integrated version of the rhamnose promoter directly controlling the expression of T7 RNA polymerase (Giacalone, Gentile et al. 2006).

Because B strains and K-12 *E. coli* strains present different metabolic and genetic backgrounds, it is expected to find substantial differences in their protein expression profiles. This is why combinatorial and high throughput screenings are usually needed to identify the best expression conditions for a certain gene and this is one of the key topics covered in this thesis.

3. Membrane proteins

3.1. General features of membrane proteins

Membranes provide compartmentalization in living organisms to isolate molecular environments for optimal function. Membrane proteins play a key role in detecting outside signals from cells, allowing them to interact and respond to their environment in a specific manner (Almen, Nordstrom et al. 2009). Membrane protein coding genes represent around 20% of all genomes (Almen, Nordstrom et al. 2009). The architecture of integral membrane proteins shows common principles, likely due to the lipid environment in which they are embedded. You can separate membrane proteins in different categories; bitopic membrane proteins in which one helix connects two domains of the protein on each side of the membrane, polytopic membrane proteins consisting of several membrane-spanning fragments and polytopic β-sheet proteins (Fig. 6). Helical polytopic membranes, also kown as α-helix bundle proteins, contain several transmembrane α -helices, each of which is around 20 amino acids long with largely hydrophobic side chains. α -helices are typically depicted perpendicular to the plane of the membrane. Transmembrane helices in polytopic membrane proteins usually have the most polar face of each helix buried in their structure while the least polar face is exposed to the lipids (Park and Helms 2008).

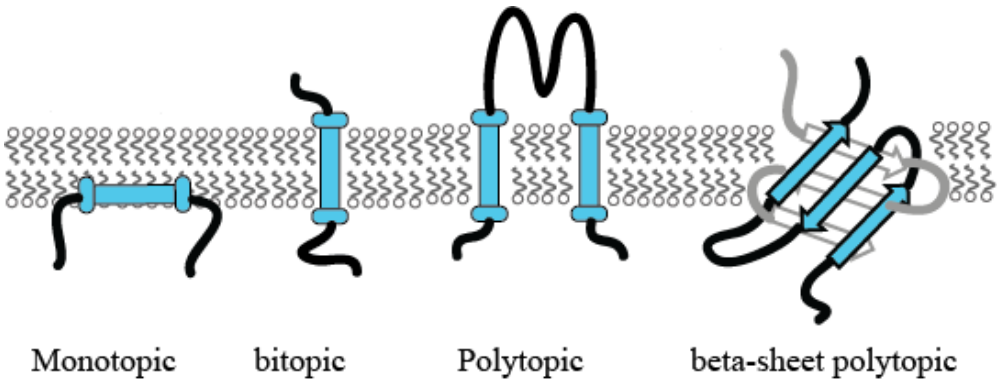


Fig.6. Drawing representation of the different types of membrane proteins

Not all membrane proteins have a clear hydrophobic profile, and are therefore difficult to identify as integral membrane proteins. These types of membrane proteins are marginally hydrophobic because they rely on the local environment to be fully integrated (De Marothy and Elofsson 2015). For example, the presence of other neighbouring helices allows for stabilization of the transmembrane domain.

The traditional fluid mosaic model describes how proteins and lipids are organized in biological membranes as a two-dimensional lipid matrix where membrane proteins diffuse freely. The model assumes that membrane proteins are interacting with lipids mainly due to hydrophobic forces (Singer and Nicolson 1972). However lateral diffusion of membrane components, protein-lipid interactions, and substrate accessibility can be essential for protein structure and function thus adding more complexity to this model. Sometimes the length of the transmembrane portions do not match the thickness of the lipid bilayer, a phenomenon termed "hydrophobic mismatch" (De Marothy and Elofsson 2015). This can cause that hydrophobic side-chains to be exposed to the aqueous environment or the hydrophilic amino acids to bury into the membrane, which in both cases is energetically unfavourable. To compensate this mismatch situation, transmembrane helices may either tilt or distort through coils in order to minimize free energy, thus preventing or causing exposure of side chains to the aqueous environment. Side chains of the transmembrane helices interact with the lipid layer and can stabilize the protein structure. These are important structural factors that may shape protein function, and should be taken into account when engineering membrane proteins in cell factories.

3.2. Translocation and insertion machinery of membrane proteins

Nearly all integral membrane proteins require the aid of protein-conducting channel proteins called translocons. A translocon can be defined as a gate in the membrane through which any protein can be transported or translocated (Holland 2004). A translocase would be an enzyme that catalyzes the actual move of the enzyme through the translocon. Key structural features of translocon complexes that mediate the translocation or insertion of membrane proteins have been solved (Holland 2004). Biochemical studies have yielded important insights into key aspects of membrane protein insertion and folding. Although there are at least two other class of protein transport across gram negative bacteria membranes, translocon-mediated transport is the most important one for protein insertion into the membrane compartment particularly for inner membrane proteins (IMPs) (Luirink, Yu et al. 2012). The other two classes of protein translocation systems across membranes are the so called auto-transporters involved in the transport of proteins through the outer membrane. And transenvelop channels that transport unfolded polypeptides beyond the outer membrane (Holland 2004).

The essential protein translocation system is known as the Sec translocon, a well conserved three-subunit membrane complex. In *E. coli* the Sec translocon comprises two heterotrimeric integral membrane complexes: SecY, SecE, and SecG (SecYEG) and SecD, SecF, and YajC (SecDF/YajC) (Mao, Cheadle et al. 2013). The Sec system translocates proteins before they fold into the tertiary structure. The system includes cytosolic factors such as SecA and SecB, which act as stabilizing chaperones to maintain newly

synthesized proteins (post-translationally) in a state compatible with the translocation. In eukaryotes this system is found in the endoplasmic reticulum (ER), and named Sec61. Several partner proteins such as SRP (Signal Recognition Particle) and YidC participate in the co-translational translocation of the polypeptide chains through what is known as the SRP-dependent pathway. The SRP binds to highly hydrophobic amino acids in the N-terminal of the nascent chain and delivers it to the SecYEG translocon. The SRP receptor called FtsY it is also required to release the nascent polypeptide chain in the membrane. The nascent chain is then pulled by the SecYEG translocon, a process that consumes energy in form of GTP (Fig. 7). (Valent, Scotti et al. 1998).

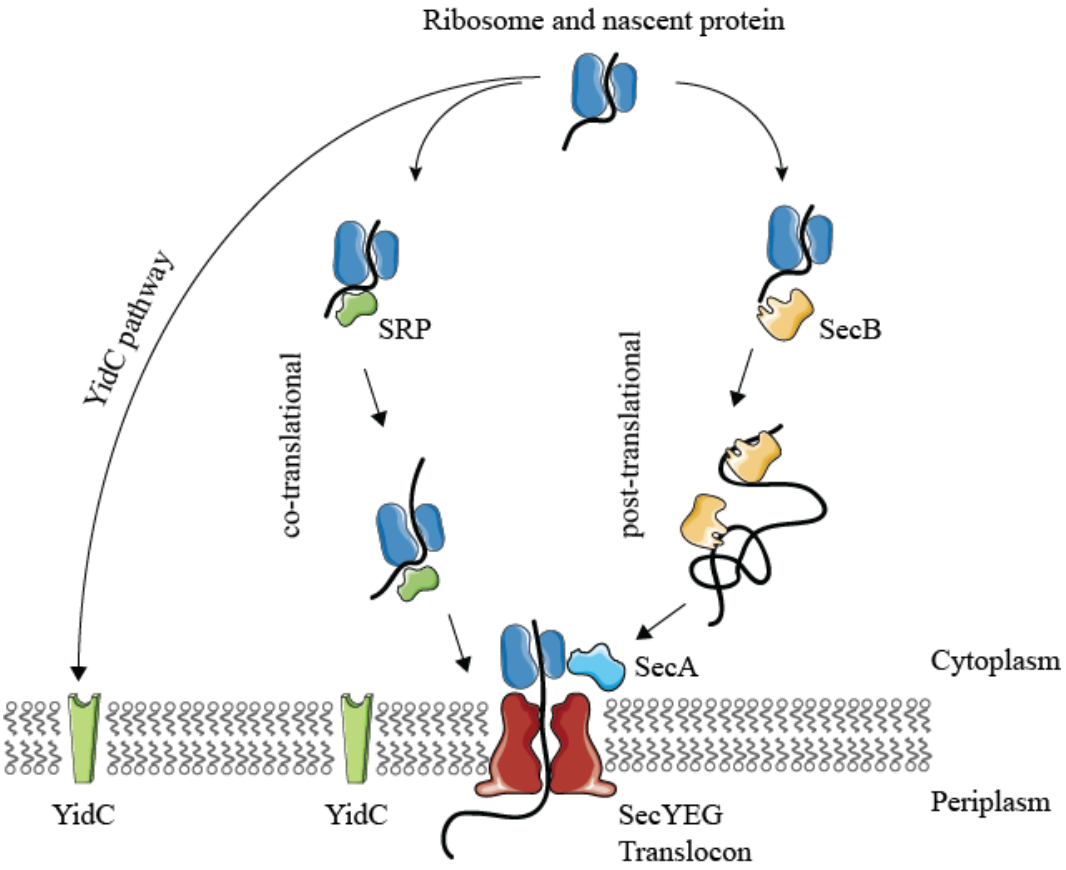


Fig. 7. Membrane protein biogenesis mechanisms of *E. coli*. Adapted from (Luirink, Yu et al. 2012).

In Bacteria the SRP-dependent pathway targets mainly membrane proteins while in eukaryotes this pathway is employed by secretory proteins as well (Denks, Vogt et al. 2014). How a cargo protein is targeted to the SRP pathway? The answer relies on the hydrophobicity of the N-terminus of the target protein, which often is given by N-terminal so-called signal peptides or short sequences with a markedly hydrophobic core that are recognized by the SRP. Based on the presence of cleavage motifs, and hydrophobicity, signal peptides can be accurately predicted, for example using SignalP (www.cbs.dtu.dk/services/SignalP) (Nielsen, Engelbrecht et al. 1997). Other prediction tools like the

Signal-BLAST are also available for the same purpose (http://sigpep.services.came.sbg.ac.at/signalblast.html) (Frank and Sippl 2008). In this thesis we have used the SignalP prediction tool to search for signal peptide fusions that can be used to facilitate better production of membrane proteins (Fig.8).

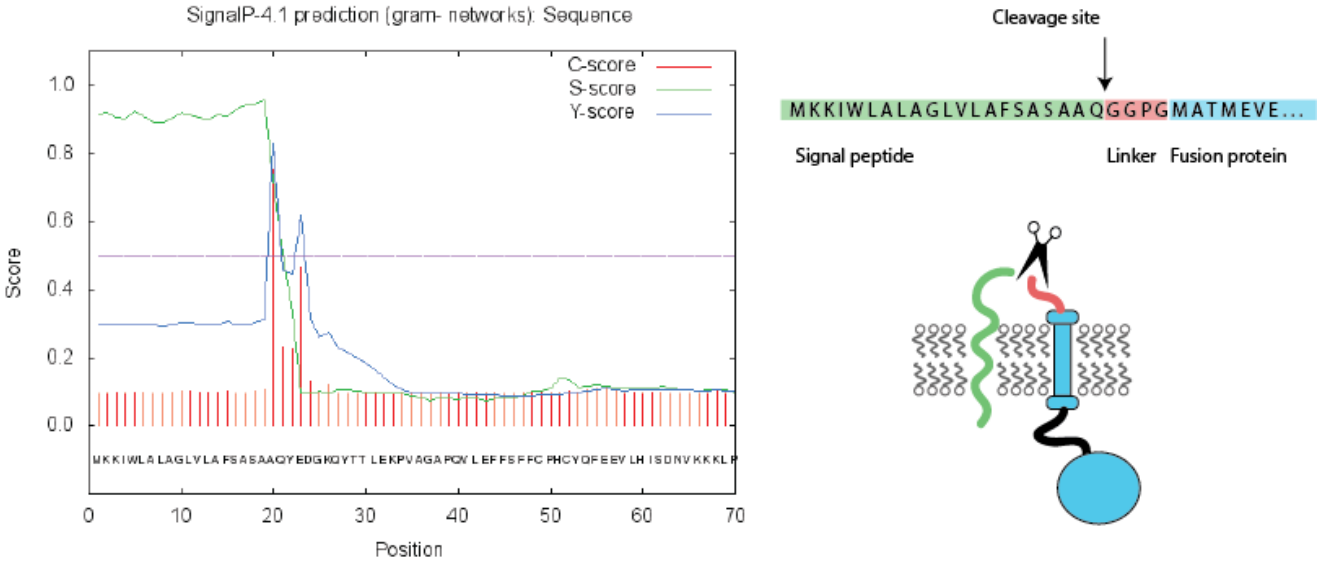
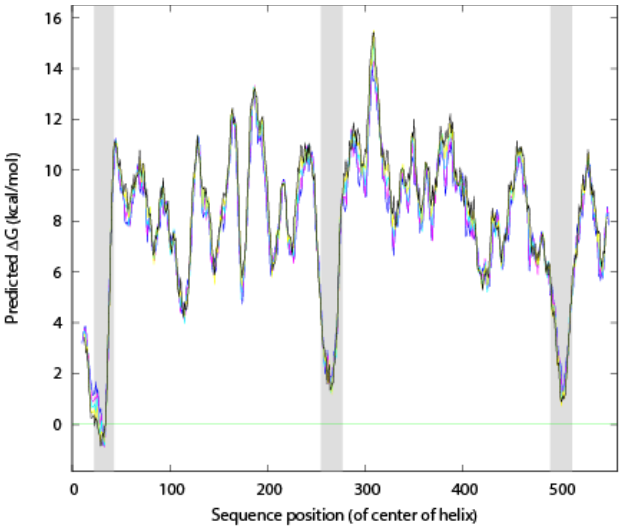
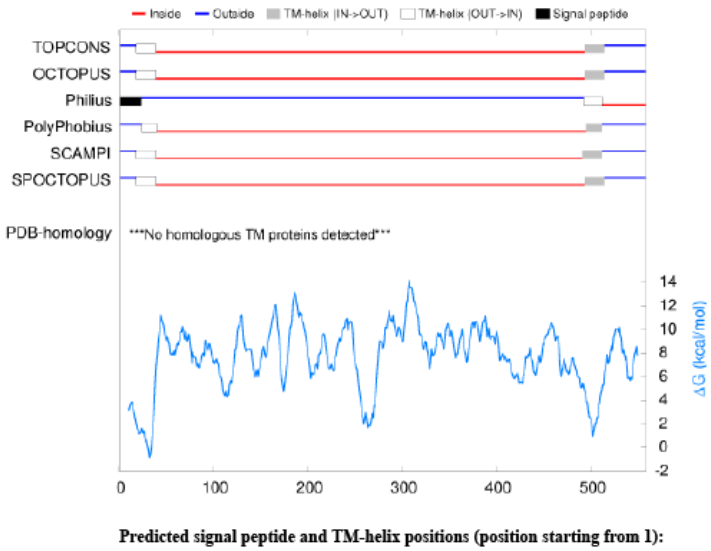


Fig. 8. On the left side a SignalP 4.0 server prediction of the periplasmic *E. coli* protein Disfulfide Oxidase (DsbA). On the right side the sequence of the predicted signal peptide and the cleavage site shown together with a schematic representation of a resulting fusion construct.

Besides the SRP-dependent pathway, it has been also observed that certain proteins can be inserted in the membrane aided by YidC, a small insertase (Luirink, Yu et al. 2012). YidC contains 5 transmembrane domains and one cytoplasmic domain, and it is speculated to work as dimer. It has been identified as an indispensable element for assisting insertion and folding of membrane proteins (Samuelson, Chen et al. 2000). It has been shown that inhibition of YidC affects the insertion of most of the Sec-dependent proteins (Samuelson, Chen et al. 2000). Interestingly, exported proteins are no so much affected by this inhibition. The cytoplasmic domain of YidC does not seem essential for its function suggesting that the key relies on the transmembrane domains (Luirink, Yu et al. 2012). On the other hand SecB is a cytoplasmic chaperone that binds mature domains of proteins and targets them to the membrane. When SecA recognizes SecB and the preprotein, is activated and binds to the membrane-embedded translocon. SecB is then released from the pre-protein as the SecA mediates posttranslationaltranslocation through the SecYEG translocon with the requirement of ATP (Valent, Scotti et al. 1998).

Behind the translocation machinery, biophysical factors govern protein insertion and translocation. Each amino acid has been shown to contribute differently to membrane insertion (Park and Helms 2008). A phenomenon known as the positive-inside rule comes from the observation that most transmembrane segments display more positive amino acids on the cytoplasmic (in-) side (Heijne 1986). There is a relationship between the free energy (ΔG) of the amino acids conforming a transmembrane domain and insertion efficiency in the membrane (Hessa, Kim et al. 2005, Hessa, Meindl-Beinker et al. 2007). However, when the length of the transmembrane segments exceeds the thickness of the membrane hydrophobic mismatch may occur. In general transmembrane segments (TM) display similar characteristics, having central hydrophobic amino acids flanked by positively charge residues on one side. Although in some cases TMs display a positive ΔG, in the context of multi-spanning membrane proteins, the interaction with other TMs influence the insertion efficiency. These biophysical parameters can be used to accurately predict the likelihood of a segment being inserted into the membrane. For these several software have been implemented and can be found online, such as the ΔG predictor (www.dgpred.cbr.su.se) or TOPCONS (Park and Helms 2008, Tsirigos, Peters et al. 2015). These predictions tools have been used in this thesis in order to engineer, substitute and fuse new membrane protein transmembrane segments (Fig. 9). Taking into account all these factors one can rationally engineer membrane proteins or even convert a soluble protein into a membrane protein (Norholm, Cunningham et al. 2011).





Predicted TM helices:

Position	*Length*	*Predicted ΔG*	*Sequence*
23-42	20	-0.859	AILKLLLFVVTLSYLARALR
255-276	22	1.240	AVFTSLGLLYAFCVSDYLPWLR
490-511	22	0.721	RRGCIAASLGTAMSVMLFGRLL

Predicted signal peptide and TM-helix positions (position starting from 1):

TOPCONS	TM1: 18-38,	TM2: 493-513
OCTOPUS	TM1: 18-38,	TM2: 493-513
Philius	SP: 1-24,	TM1: 492-511
PolyPhobius	TM1: 24-39,	TM2: 495-511
SCAMPI	TM1: 18-38,	TM2: 491-511
SPOCTOPUS	TM1: 18-38,	TM2: 493-513

Fig. 9. Predicted transmembrane segments of the Cytochrome P450 CYP79A1. On the right side, ΔG predictor output. On the right side TOPCONS prediction of the same protein sequence including the orientation in blue and red color. The most probable transmembrane segment is highlighted in green.

3.3. Techniques to study membrane proteins

Membrane proteins are difficult to work with due to the nature of the environment in which they are embedded. Popular models for study functionality and structure of membrane proteins include whole cells, liposomes, microsomes, spheroplasts, nanodiscs, SMALPs and a handful of other systems comprising components in which proteins are packed in a lipid-like environment. These systems aim to isolate membrane proteins in lipid-like but more soluble state so that they can be easily studied.

3.3.1 Topology

In gram negative bacteria, such as *E. coli*, the cytoplasm and the periplasm are separated by a membrane. These two compartments provide different chemical environments, thus active cytoplasmic proteins might not be active in the periplasm and vice versa. For example, the periplasm displays a reducing environment optimal for disulfide bond formation which is critical for the function of some proteins (de Marco 2009). Because membranes represent physical boundaries to separate compartments, membrane topology is an important aspect that determines protein functionality. In the context of membrane proteins, topology is defined as the orientation and number of the transmembrane spanning regions with respect to the inner and outer compartment (von Heijne 2006). If a protein is inserted in the membrane in a way that the carboxyl terminal is exposed to the cytoplasmic side of *E. coli*, this protein has a C_{in} topology. Similarly if the N-terminus of the protein remains exposed to the periplasmic side, this would have an N_{out} topology. Although prediction tools have been mentioned in the previous section, experimental validation of membrane protein topology is often required. Some of the common techniques to determine protein topology are presented here:

- Compartment-specific enzyme activity: A smart way to determine whether a certain protein domain of a protein lies on the periplasmic or the cytoplasmic side, is to couple an enzyme whose activity is optimal only in one cellular compartment. The topology of the inner membrane proteome of *E. coli* was determined with this approach by Daley and co-workers (Daley, Rapp et al. 2005). For example membrane proteins can be tagged on the C-terminus with either Green Fluorescent Protein (GFP) that is only active on the cytoplasmic side or Alkaline Phosphatase (PhoA) which is only active in the periplasm (Fig. 10). The oxidant conditions of the periplasm seem to produce intra or inter-transmolecular disulphide bonds in the GFP moiety. This reversibly inactivates GFP in the periplasmic compartment, and makes it a good subcellular localization reporter (Feilmeier, Iseminger et al. 2000). In contrast, PhoA is only active in the

periplasm as it contains four cysteines residues that must be oxidized for the enzyme to be active (Belin 2010). This way the protein topology can be studied *in vivo* and the abundance of a certain protein estimated with the help of the reporter proteins. GFP fusions are used in this thesis to determine not only the right topology of heterologously expressed membrane proteins but also proper folding of the same. This platform has been also developed in the yeast *Saccharomyces cerevisiae*, and allows for identification of properly folded eukaryotic membrane proteins (Drew, Newstead et al. 2008).

- Protease accessibility: Often inter-transmembrane loops may not be large enough to fuse a reporter, or the reporter may not be active. In these cases, radioactively labelled or tagged membrane proteins embedded in some of the model lipid systems mentioned above can be cleaved with a specific or broad protease generating fragments of different sizes (Fig 11). The result is analyzed by traditional SDS-PAGE to confirm expected fragment sizes. This approach together with fusion reporters has been used to determine the topology of the Oxa1P homologue in *E. coli* (Saaf, Monne et al. 1998).

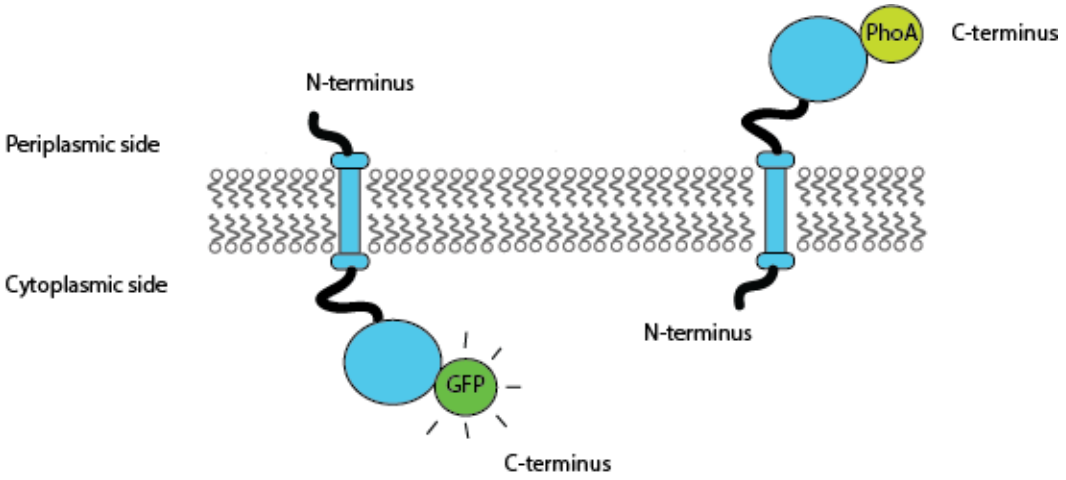


Fig. 10 Membrane protein topology mapping using fusion reporters. Fusion reporter GFP (left) is active in the cytoplasmic side whereas Alkaline Phosphatase (PhoA) on the right, is only active in the periplasm.

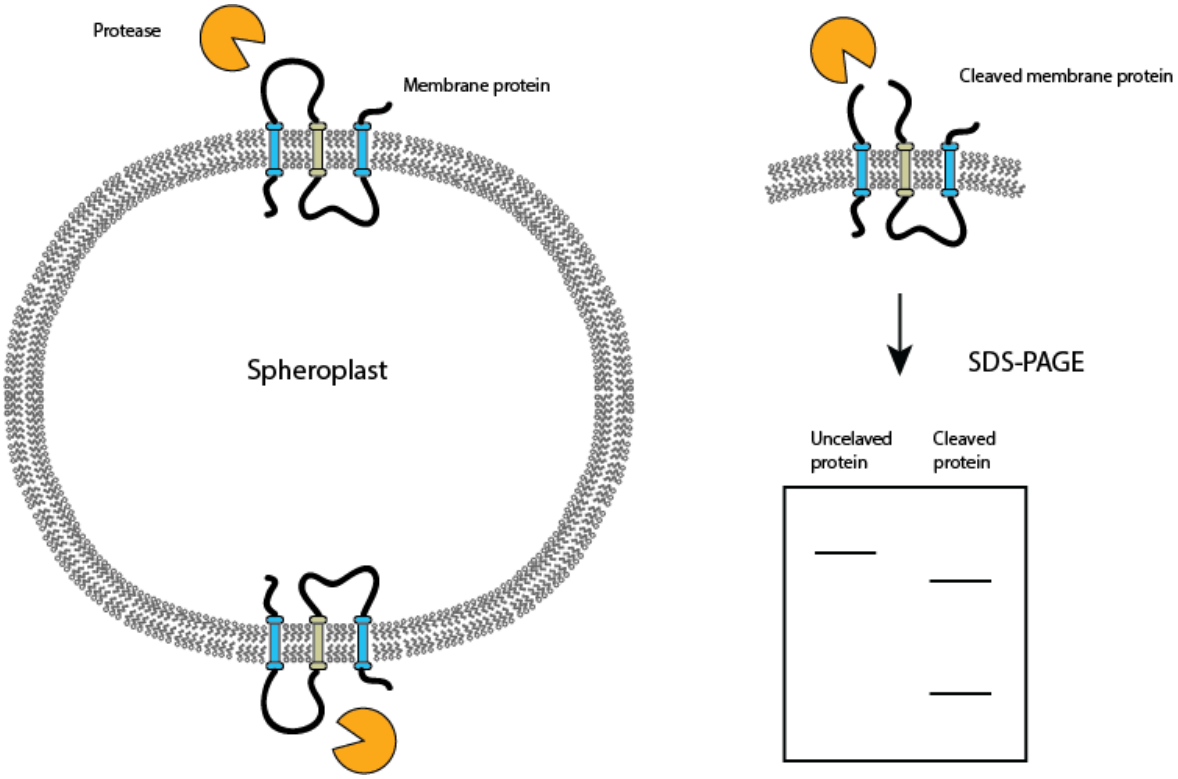


Fig. 11. Schematic representation of a protease accessibility assay. A lipid system such as spheroblasts (a bacterial cell without the cell wall) where the membrane proteins are located, is digested with a protease that access external loops. The resulting products of the digestion can be analyzed by SDS-PAGE.

- Glycosylation motifs: An effective approach to mapping the topology is carried out using an oligosaccharyl transferase, which catalyzes addition of oligosaccharides to the amino group of asparagine residues within the consensus sequence Asn-X-Thr/Ser. N-glycosylation is a common feature of eukaryotic membrane proteins, and the consensus sequence is usually found in the largest luminal exposed loops of the protein. Since modification of the glycosylation site occurs in a com-partment-specific manner, the presence of glycosylation provides information for topological assignment (Bogdanov, Zhang et al. 2005). These motifs can be introduced in membrane protein loops by standard recombinant DNA technologies, and expressed in vitro in the presence of canine pancreas microsomes (Goldman and Blobel 1981, Hessa, Kim et al. 2005). Glycosylation occurs co-translationally, thus glycosylated loops indicate that a given protein segment has crossed the ER membrane, and the topology can be determined by SDS-PAGE based on the different molecular size of the glycosylated and unglycosylated protein.

3.3.2 Structure

Crystallography: X-ray crystallography is one of the most popular techniques for structure determination of proteins. Increasingly, researchers interested in all branches of the biological sciences require structural information to answer relevant questions. Protein crystallization is a process of testing a large number of possible crystallization reagents. Once initial crystallization conditions are found, further optimization is usually necessary to obtain well-diffracting crystals. Usually the screening starts using 96-well plates. In the past 10 years, the volumes required for these crystallization experiments have been substantially reduced. Although there are more than 30 different 96-well sparse matrix screening systems, membrane proteins are difficult to crystallize.

Screening the detergent is a critical step in crystallization because they are utilized for solubilizing membrane proteins, thereby creating a similar environment to the natural lipid in which they are usually embedded. Detergents are amphipathic molecules, consisting of a polar head group and a hydrophobic tail. Their display unique features, for example they can organize in spontaneously in spherical structures called micelles (Seddon, Curnow et al. 2004). The minimum concentration of detergent needed to cluster in micelles is called Critical Micellar Concentration (CMC). Some membrane proteins may respond better to detergents with high CMCs, while others are more stable at lower CMCs. This is why screening different buffer and detergents is often necessary to determine the best combination that gives a monodisperse form, namely, micelles of uniform size in the solution (Bogdanov, Zhang et al. 2005). Once well-diffracting crystals are obtained data from beam diffraction can be obtained. Due to the large number of conditions to optimize, high protein amounts are usually required.

NMR (Nuclear Magnetic Resonance) is another popular way to elucidate protein structures based on the unique magnetic properties of nuclei particles, typically from ^{2}H, ^{15}N and ^{13}C isotopes. When a magnetic field is applied at a certain frequency to the sample, the atomic nuclei particles absorb and emit electromagnetic radiation in different ranges. The magnetization of the nuclei can also be transferred to the surrounding atoms producing specific radiation patterns depending on the chemical bond. This output can be converted into structural information creating a connection map of the atoms conforming the protein. Unfortunately, multidimensional NMR still requires high amounts of solubilized and stable protein because they have to be labelled with multiple isotopes, being protein expression the main bottleneck. Typically NMR methods have been limited to small proteins, and the denaturation process of proteins greatly reduces the resolution of this technique. However, recently developed NMR methods may in the future allow for elucidation of larger proteins thanks to optimized spectroscopic techniques

(Fernandez and Wuthrich 2003). In addition to the challenges associated to protein size and expression levels, not all detergents are suitable for NMR applications. Because of the interactions between different detergents and the amino acid side chains located in the flexible loops or other intra-molecular backbone regions, most membrane proteins cannot be fully characterized using NMR (Fernandez and Wuthrich 2003). All these factors have delayed the use of NMR as a preferred method for structural determination of membrane proteins.

Microscopy techniques have also received a lot of attention lately.Unlike crystallography electron microscopy allows for study membrane proteins in two-dimensional (2D) crystals with resolutions that can reach 3 Å or better. 2D membrane protein crystals are frequently grown easier than 3D crystals and offer a more native environment than most 3D (Fig. 12). Recent technological advances in the cryo-electron microscopy (cryo-EM) equipment has led to the expansion of this method in the membrane protein field. For example, the atomic resolution structures of bacteriorhodopsin and several other membrane proteins have been determined using cryo-EM (Fujiyoshi 2011). In addition, a large number of membrane protein structures were studied at a slightly lower resolution, whereby at least secondary structure motifs could be identified (Fujiyoshi 2011) Crystallographic and NMR determination of membrane proteins requires a large amount of stable protein in a detergent for at least few days during measurement. Sometimes this is not possible, especially for many eukaryotic membrane proteins. In contrast, in cryo-EM, membrane proteins are quickly frozen and the image of the electrons crossing the sample captured as single molecules in a thin film of buffer (Vinothkumar 2015). By averaging a large number of particles, high-resolution structures can then be obtained with small amounts of expressed protein. This technique allows for capturing multiple structural states of the same protein in the solution, which can be computationally separated. Finally it is worth mentioning other new microscopy and spectroscopic techniques such as Atomic Force Microscopy (AFM) that are promising in the field of membrane proteins. AFM gives the chance to mechanically manipulate membrane proteins in its native state and opens the door to discover new properties of membrane proteins previously not considered (Muller and Engel 2007).

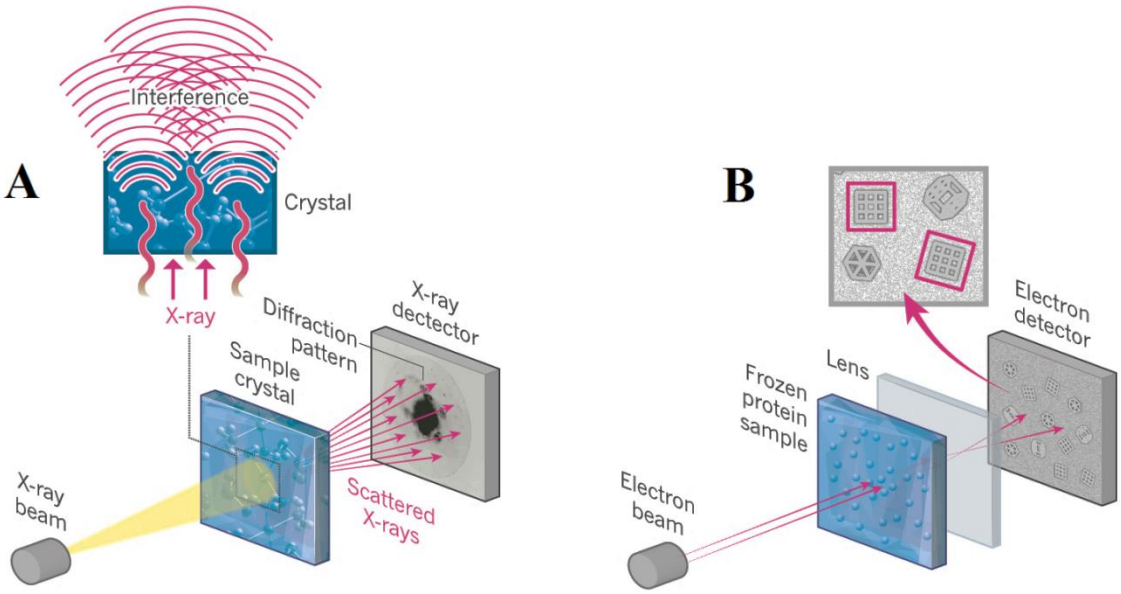


Fig. 12. X-ray crystallography is based in a high intensity radiation beam penetrating protein crystals creating a unique scattering pattern (**A**). In electron microscopy an electron beam passes through a frozen protein solution and scattered electrons cross a lens creating a magnified image (**B**). Image copied from *Nature*©.

A part from experimental methods, there are alternative ways to deduce structural features of membrane proteins using computational tools. One of them is homology-based modelling, which relies on similarities between unknown and known structures. By exploiting structural information from the known structures, the new structure can be approximated computationally. This approach may become even more feasible in the future thanks to initiatives such as the structural genomics project, which aims to elucidate at least one structure of each protein family (Burley, Almo et al. 1999). Homology modeling is typically carried out in four steps: **1)** Find known structures similar to the target sequence. **2)** Sequence alignment. **3)** Build a protein model. And **4)** assess the model.

Known 3D structures, so-called templates, can be found in repositories such as the Protein Data Bank (PDB). For modelling, sequence comparison methods or sequence-structure threading methods are frequently used. The latter can sometimes reveal more distant relationships than purely sequence-based methods as we will point out in chapter 3 of this thesis. There cytochromes P450 sequence alignments are largely performed with scoring matrices called BLOSUM. The Blosum62 matrix detects distant relationships between the target sequence and

all sequences of a certain repository producing an alignment of diverged proteins more accurate for sequences with known three-dimensional structures. The accuracy of a homology model is related to the percentage sequence identity between the structural and sequence similarity of the sequences. The error is measured as the root mean square (RMS) for the main-chain atoms. Homology models can be sorted into three different categories: **1)** High accurate models with more than 50% sequence identity to their templates, comparable to the accuracy of a medium-resolution nuclear magnetic resonance (NMR) structure or a low-resolution x-ray structure. Here errors are mostly located in side-chain packing, small distortions of the core main-chain regions, and occasionally larger errors occur in loops **2)** Medium accurate models with 30 to 50% sequence identity. Here 90% are errors in the range of 1.5 Å located over the main chain. Side chains, core distortion, and loop modelling errors tend to be more frequent. **3)** Low accurate models with less than 30% sequence identity (Baker and Sali 2001).

There is a myriad of applications of homology or comparative modelling. For example, validation of functional predictions that have been based purely on a sequence similarity. Ligand binding is linked to the structure of the binding pocket and difficult to deduce from the sequence composition so a homology model becomes very handy for validation. Moreover the size of a ligand may very well be predicted from the volume of the binding pocket, especially with high accurate homology models. For these reasons, modelling is very popular in drug discovery, as it feeds on known structures to generate models of, for example, unknown receptors that are drug targets (Baker and Sali 2001). The ultimate goal of generating models in this case is to perform docking studies. Docking studies involve computational calculations to find possible thermodynamically relevant states of a ligand in the binding site. Several binding modes can be generated and evaluated using scoring systems. This technique has been widely used in drug discovery for virtual screening of multiple ligands, and large compound libraries without the need for initial experimental work. However, in search for new drugs from natural environments homology models and docking are becoming powerful tools for discovery of both intermediate chemicals and pathways leading to a certain compound. For example, with this approach, new intermediates of the glycolytic pathway have been discovered in *E. coli* (Zhao, Kumar et al. 2013). The promising results have inspired us to develop more accurate methods for homology models that may be useful in metabolic engineering applications of cytochromes P450. The results of our study published in the third chapter of this thesis suggest that homology-model-based methods may be utilized to decipher key enzymatic mechanisms of structurally unknown plant P450s.

3.4. Challenges in membrane protein production

The nature of lipid bilayers and biological membranes represent a technical challenge to study proteins embedded in. Their surface is relatively hydrophobic and they have to be extracted from the cell membrane using different detergents. For this reason membrane protein characterization present challenges at many levels, including expression, solubilization, purification, crystallization, data collection from all crystals generated in the screening and the final structure solution (Carpenter, Beis et al. 2008). One of the major bottlenecks in the whole process is undoubtedly protein expression as it involves testing of a large number of conditions to find the best one for each protein. Before some of the latest advancements in the membrane protein overexpression technologies another of the concerns was to predict or assay whether a certain protein would end up in the membrane or in inclusion bodies.

Nowadays *E. coli* is regarded as one of the preferred hosts for membrane protein expression as it represents a quick, relatively inexpensive and easy to use system enabling many constructs to be screened quickly. However, for some heterologous proteins there is no alternative but the use eukaryotic hosts such as the yeast *S. cerevisiae*. First, membrane proteins have to be targeted to the cell membrane so that they are most likely to fold correctly. Several tricks have to be used for targeting hetereologous proteins through the Sec translocon in order to be inserted in the membrane, for example signal peptides or engineered N-terminus sequences (Schierle, Berkmen et al. 2003).

Second, as mentioned in the section 3.3.2 the choice of detergent to extract proteins from the membrane greatly affects the yield of functional protein. Often several detergents must be screened to identify the detergent that extracts the largest quantity of soluble, active, homogeneous, stable protein, provided that the cost of the detergent is not limiting. One of the most effective detergents to comply with these requirements is dodecyl-maltoside (DDM) as it is relatively cheap and can give stable membrane proteins (Carpenter, Beis et al. 2008). Similarly plants cytochromes P450 are a class of membrane-bound proteins, that display some of the same challenges in expression and characterization due to its hydrophobic nature. All these factors together make the overexpression of membrane proteins more an art than a predictable technology. However, recent advancements have contributed to more streamlined procedures to tackle membrane protein overexpression, particularly in *E. coli*. Similar platforms have been also developed in the yeast *S.cerevisiae* by Drew and co-workers, giving a fairly high chance to obtain a desired protein with this host (Drew, Newstead et al. 2008). When it comes to expression of eukaryotic proteins, yeast offers obvious advantages with respect to *E. coli*. For example, many eukaryotic proteins exhibit post-translational modifications such as glycosylations and yeast contain a set of oligosaccharyltransferases capable of transferring sugars to proteins. Yeast has also a different lipid composition than bacterial membranes, and the eukaryotic translocation machinery, which specifically recognizes eukaryotic signal sequences. Although expression platforms are comparable to the ones

developed in yeast, this thesis focuses on membrane protein technologies to solve plant cytochromes P450 expression bottlenecks in *E. coli*.

3.5. *Escherichia coli* as membrane protein production and screening host

In the section 2.3 some of the most widely used *E. coli* expression strains are presented. As stated in the same section one of the main advantages of *E. coli* is the simplicity of its manipulation, and the doubling time. Moreover *E. coli* grows to higher cell densities; therefore scale-up procedures are usually faster. Unlike eukaryotes, *E. coli* does not contain subcellular organelles or endomembranes so all membrane proteins must be localized in the cytoplasmic or outer membrane. Another virtue of this microbe is the availability of genetic tools and a large mutant collection that in case an endogenous gene interferes with the heterologous product, the background can be greatly reduced. Finally transformation of DNA into *E. coli* cells is also a very quick protocol, which enables for high-throughput experiments (Rosano and Ceccarelli 2014).

Given the mentioned challenges in membrane protein production, several solutions have been developed thanks to the advantages of this host. In this thesis we discuss two particular technologies to address the bottlenecks abovementioned: **1)** Strains able to cope with expression of toxic genes. Particularly the use of the Lemo21(DE3) strain, and the KRX strain both of which bear a titratable rhamnose promoter to allow for gradual gene expression avoiding saturation of the Sec translocon complex and proper co-translational folding of the target protein (see section 2.3). **2)** Screening platforms that discriminate properly folded membrane proteins facing the right cell compartment. More specifically a C-terminus Histidine-tagged-GFP reporter serves to quickly asses integration and folding of overexpressed membrane proteins. This fluorescence-based optimization scheme not only enables estimation of protein production levels but also capture hetereologous protein by the histidine tail, facilitating further solubilization screening and purification of the membrane protein fusion (Drew, Lerch et al. 2006).

3.6. Applications of membrane proteins in metabolic and protein engineering

3.6.1. Metabolite transport

The main means by which small molecules cross biological cell membranes is through proteinaceous transporter molecules, and this gives the cells means by which to control the process (Kell, Swainston et al. 2015). An important application of membrane proteins in cell factories and metabolic engineering is the use of influx and efflux transporters. For example introducing new transporters, can give rise to higher toxicity tolerance or facilitate secretion of

the product of interest to avoid its degradation. For this reason there is a great interest in balancing an effective efflux of newly synthesized compounds. More information about structure-function relationship is also in great need in order to select for the best transporters (Kell, Swainston et al. 2015). Transporters can also provide a solution to osmotic stressed caused by the accumulation of high titers of the product inside the cells, acting as a pressure valve to relief cells.

3.6.2. Membrane scaffolds

In nature, an important number of biochemical reactions are not catalyzed by isolated enzymes but by multienzyme complexes (Proschel, Detsch et al. 2015). These so-called metabolons represent micro-compartments to balance metabolic pathways thereby co-localizing their enzymes in a specific location, for example the Endoplasmic Reticulum (ER) membrane (Jorgensen, Rasmussen et al. 2005). Membrane domains may act as scaffolds of a particular enzyme, anchoring it to the membrane compartment where several other enzymes are brought into close proximity in a semi 2d-space. This may facilitate or accelerate channeling of the intermediates through a metabolic pathway. This feature becomes very useful when intermediates of the reaction are unstable because they can be quickly converted by the neighboring enzymes. Consequently scaffolding aims to co-localize the components of a metabolic pathway in order to provide spatial and temporal control of molecules (Proschel, Detsch et al. 2015). Endogenous *E. coli* membrane proteins known to interact in the inner membrane, such as the recently discovered periplasmic chaperone YfgM and PpiD (Fig. 13) (Gotzke, Muheim et al. 2015) could be used to create artificial metabolic channels.

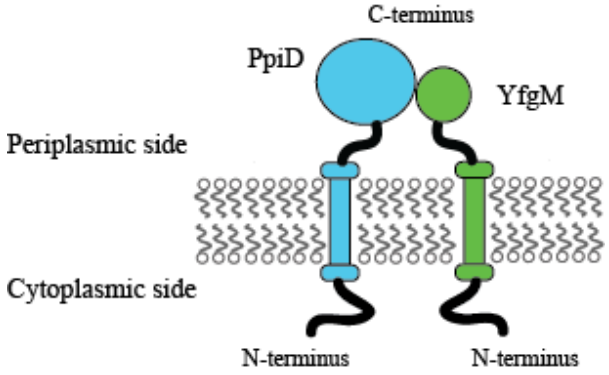


Fig. 13. Schematic representation of YfgM and PpiD interaction in the inner membrane of *E. coli*.

Many natural products such as retinol, show high hydrophobicity and tend to stick to membrane compartments (Lhor and Salesse 2014). In this context membrane scaffolds may be

advantageous because they would co-localize with the environment where the substrates are embedded in, facilitating catalysis. Furthermore the presence of the biocatalysts in the membranes may help to relief the physiological stress that represents an organic compound within the lipid bilayer (Murinova and Dercova 2014). In the current thesis we apply some of these strategies to cytochromes P450 enzymes, proteins naturally found in associated to the ER membrane. Many substrates of this enzyme family are indeed hydrophobic as they are derived from isoprenoid backbones which are built up from aliphatic chains (Hill and Connolly 2015).

3.6.3. Surface and periplasmic display technologies

Cell surface display comprises a variety of techniques that allows for expression of proteins or peptides on the surface of cells in a stable manner using the surface proteins of bacteria as fusion partners. Numerous scaffolds proteins or peptides such as OprF, OmpC, OmpX, and others, have been used to present peptides and proteins on the outer surface of *E. coli*. The scaffold protein has to be able to transport the desired passenger protein to the external surface of the cell. The size, folding efficiency, and disulfide bond number of the passenger protein can strongly influence its ability to be secreted (Mergulhao, Summers et al. 2005). Protein display has been used for a broad range of applications, such as vaccine development, peptide libraries screening, whole-cell catalysis or biosensors and environmental bioremediation. This technique is particularly interesting when for example; the substrate of a certain enzyme cannot passively diffuse through the cell membrane due to its size or chemical composition. There could be that the substrate does not enter the cell because of the lack of a specific transporter. For example, in biomass conversion of cellulose to biofuels or assembly of polymers outside the cell (Mazzoli 2012).

In this thesis we have utilized a variant of the surface display technology called Anchored Periplasmic Expression Technology (APEx), which was developed to detect high affinity antibody-epitope interactions in the periplasmic space of *E. coli* cells. In the APEx technology, the peptide from the lipoprotein NlpA of *E. coli* acts as scaffold for passenger antibodies that are translocated to the periplasm (Jeong, Seo et al. 2007). This highly expressed peptide bears a short lipid anchoring sequence (CDQSSS), which localizes to the inner leaflet of the cytoplasmic membrane and can also be used to accommodate other proteins such as cytochromes P450 in the inner membrane of *E. coli* (**Fig. 14**). This technology does not only provide a small peptide fusion for improving expression but also a way to specifically target proteins to the inner membrane of *E. coli* (Jeong, Seo et al. 2007).

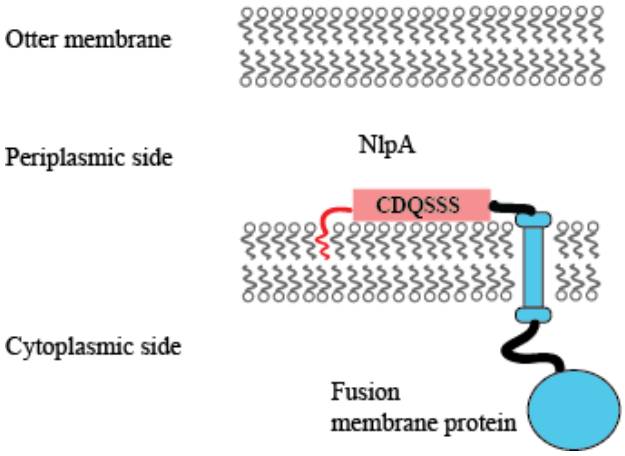


Fig. 14. Application of the APEx technology to the targeting of membrane protein fusions in *E. coli*. The small signal peptide and anchoring sequence from the lipoprotein NlpA is used to attach a protein to the inner leaflet of the cytoplasmic membrane of *E. coli*.

4. Plant cytochromes P450

4.1. General features of cytochromes P450

Cytochromes P450 (P450s) are a large family of oxidative hemoproteins (containing a porphyrin group with an iron atom) that are found in prokaryotic and eukaryotic organisms. Cytochrome P450 proteins can be classified in two groups; 1) class I (prokaryotic/mitochondrial) and class II (eukaryotic microsomes). Most prokaryotes and mitochondria have 3-component P450 systems comprising a FAD-containing flavoprotein (NAD(P)H-dependent reductase), an iron-sulphur protein and P450. In contrast eukaryotes microsomes have 2-component P450 systems composed by NADPH:P450 reductase (FAD and FMN-containing flavoprotein). P450s were first discovered in 1955 in rat liver microsomes and are characterized by showing an intense absorption peak at 450 nm wavelength in the presence of carbon monoxide. In 1987 the first X-ray structure of the camphor-metabolizing monooxygenase (P450cam) from the bacterium *Pseudomonas putida*, became a structural and mechanistic paradigm for this protein family until the P450 BM3 structure was solved in 1993 (Ravichandran, Boddupalli et al. 1993). To a large extent the knowledge about the molecular level structure-function relationships in P450s is based on studies with the P450cam system(Poulos 2003).

Importantly, eukaryotic P450s are typically bound to the ER membrane due to the presence of a transmembrane helix in their N-terminus (see Fig. 9). They are co-translationally inserted leaving the globular domain exposed to the cytoplasmic side of the ER (Monier, Van Luc et al. 1988). P450s also display a remarkably high hydrophobicity because of the presence of several other long and conserved helices further downstream in the peptide sequence (e.g. the so-called H, I, and K helices (Fig. 15)). Other P450 specific motifs are indispensable for P450 function; **1)** A heme binding domain that stabilizes the group characterized by a conserved EXXR motif at the end of the so-called K-helix. **2)** Another conserved WXXXR sequence located in the N-terminal side of the C-helix and **3)** a strictly conserved cysteine heme-coordinating amino acid located close to the C-terminal end (Sezutsu, Le Goff et al. 2013). Besides the well conserved regions, variable elements of the P450 structure determine their function. There are up to six of the so-called substrate recognition sites (SRS) that play a major role in substrate interaction, and can be identified by their position with respect to the conserved regions (Fig. 15) (Sirim, Widmann et al. 2010).

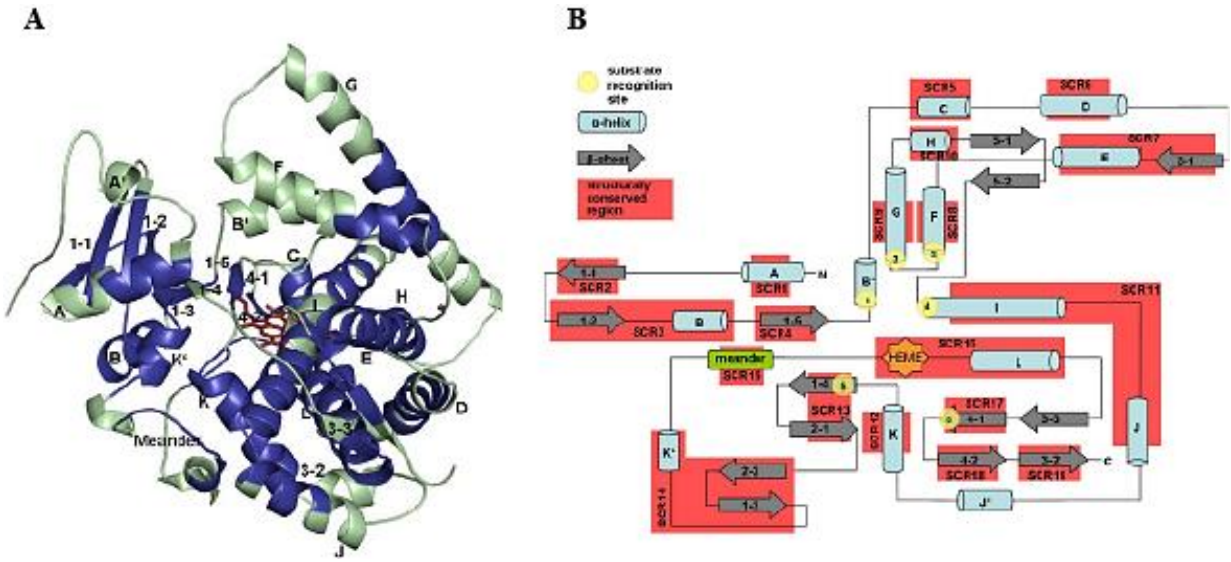


Fig. 15. Structurally conserved regions highlighted in blue over the P450 BM3 crystal structure (A). Schematic representation of the different conserved and variable motifs found along P450 structures (B). Image adapted from Demet Sirim (Sirim, Widmann et al. 2010).

Variable regions are speculated to be involved in substrate access, exit channels, substrate selectivity, specificity and stereospecificity of their substrates. However, it is protein dependent which of the SRS are involved in the different mechanisms. For this reason methods alternative to X-ray crystallography or NMR are greatly needed for elucidating structure-function relationships.

4.2. Catalytic cycle of P450s

Cytochromes P450 catalyze a wide variety of hydroxylation reactions in primary and secondary metabolism. However even nowadays, the exact nature of the active species responsible for the oxygen insertion step still remain a major point of debate. The consensus mechanism is summarized in figure 16 (Meunier, de Visser et al. 2004). Molecular orbitals in the prosthetic group govern the initial steps in the catalytic cycle of P450s, whereas the environment of the active site contribute to the formation of the catalytically active enzyme. In the second chapter of the thesis, we explore some of the reaction mechanisms of P450s related to their structural elements.

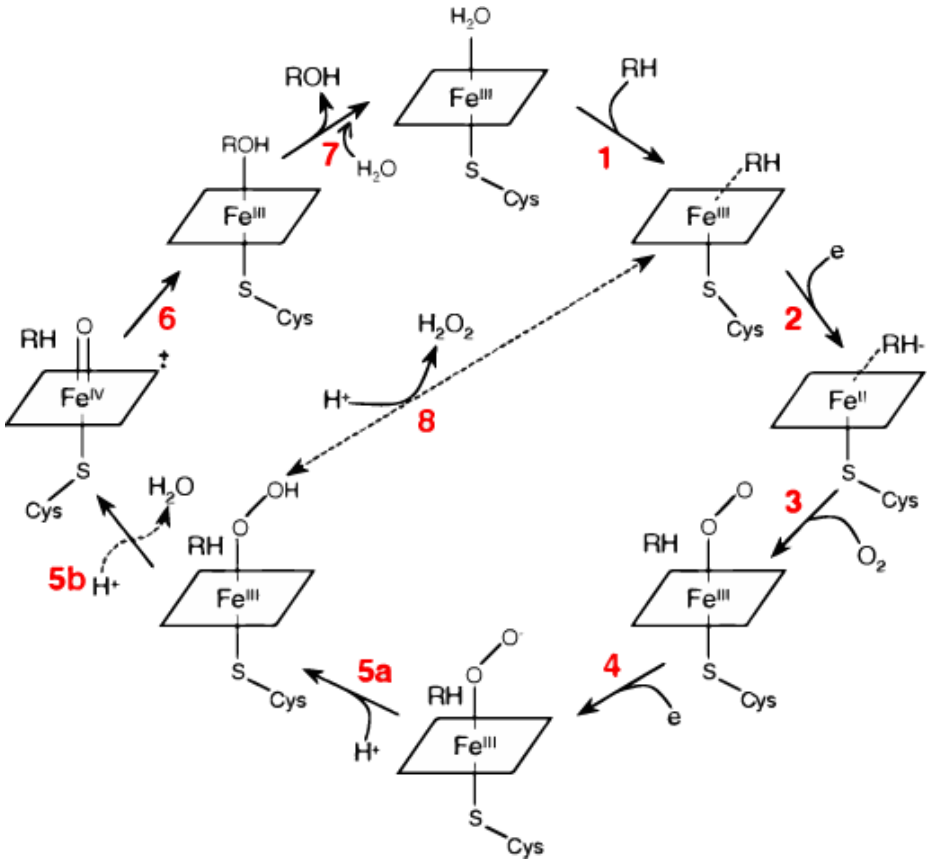


Fig. 16. Catalytic cycle of cytochromes P450. Adapted from (Shebley, Kent et al. 2009).

The catalytic cycle of P450s begins when the substrate enters into the active site displacing axial water molecules coordinated to the iron atom, an entropy driven process (ΔS) (step 1). This causes the iron atom to drop out of the plane, and the electrons of the complex to occupy a low spin state. The porphyrin ring then becomes a good electron sink, triggering the electron transfer from the redox partner reductase (step 2). Iron is reduced from Fe^{3+} to Fe^{2+}, but the electron delocalizes throughout the prosthetic group. This negative charge has a key role in the cleavage of dioxygen bond leading to iron-oxo species because electrons occupy now a high spin state in the $d_{x^2-y^2}$ orbitals. Oxygen enters the active site and is coordinated by the Fe^{2+}. One electron from the Fe^{2+} center and one from the oxygen pair subsequently create an Fe^{3+}-oxygen bond (step 3). The Fe^{3+}-oxygen complex is a nucleophile and further dissociates into Fe^{3+} and a superoxide anion (step 4). Next, a second and rate limiting electron transfer step occurs

, generating a Fe^{3+}–peroxo complex. At this point the amino acid residues of the active site rapidly mediate stabilization and protonation of the negatively charged iron-peroxo intermediate becoming a strong nucleophile (step 5a). In particular an aspartic acid residue in position 251 and threonine 252 of the P450cam has been observed to play a major role in the protonation delivery pathway to the

iron-peroxo intermediate (Fig. 17). There are experimental evidence of the existence of an hydrogen bond and water molecules network that together with other positively charge amino acids acts to stabilize dioxo species (Meunier, de Visser et al. 2004).

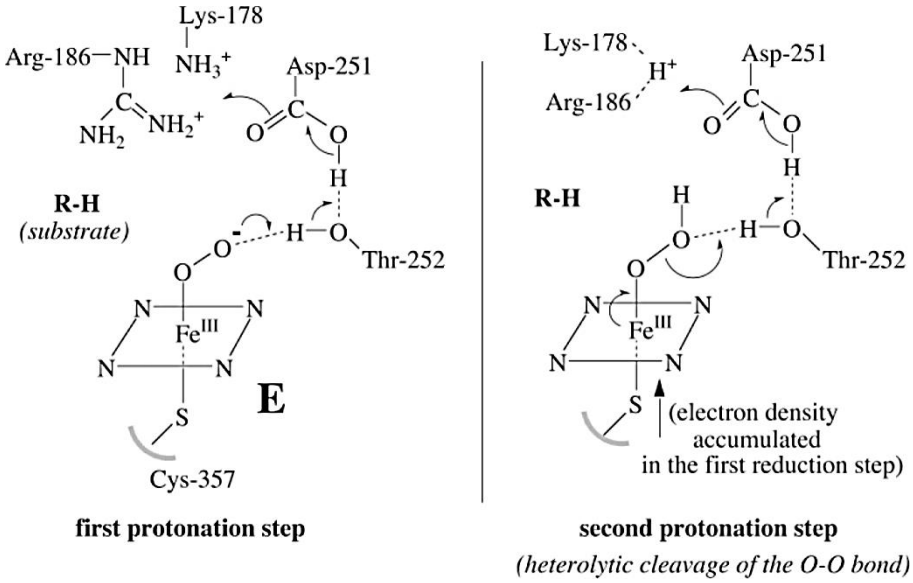


Fig. 17. Theoretical proton delivery mechanism with the participation of Aspartate 251 and Threonine 252 in the P450cam. Copied from (Meunier, de Visser et al. 2004).

The protonated Fe^{3+}–OOH stays trapped by the threonine 252 in the P450cam model an undergoes a second protonation round resulting in the splitting of a water molecule, namely heterocyclic cleavage of molecular oxygen, and generation of the so-called compound I (step 5b). The cysteine's sulfur group *p* orbital greatly contributes to the formation of this intermediate by concentrating a negative charge density. The formation of compound 1, changes the oxidation state of Fe^{3+} to Fe^{4+} with the formation of iron-oxo species Fe=O leaving molecular orbitals π_{xz} , π_{yz} and a_{2u} occupied. Consequently, the iron-oxo intermediate becomes a good electrophile capable to subtract electrons from the substrate (Fig. 18).

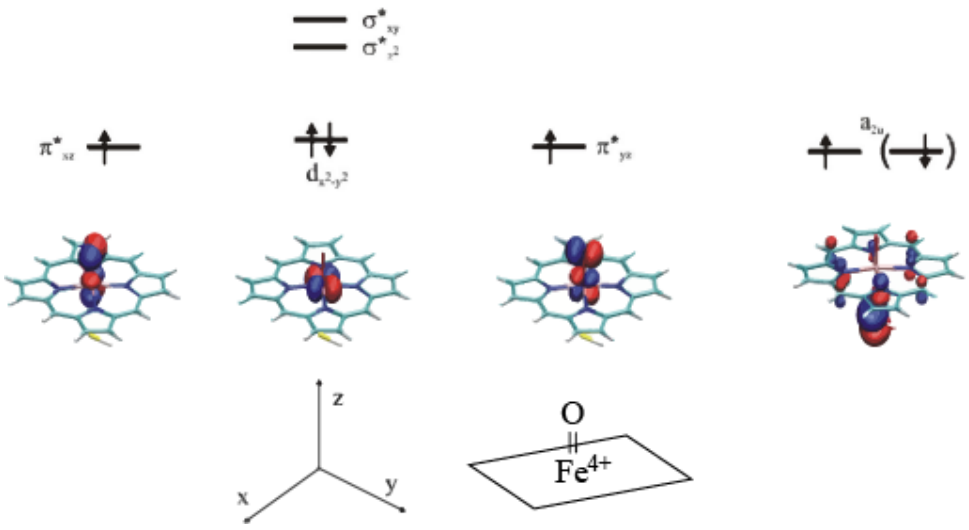


Fig. 18. Molecular orbitals occupied in the compound 1 of the P450 catalytic cycle. Adapted from (Naray-Szabo, Olah et al. 2013).

The last step of the cycle involves the oxidation of the substrate through compound 1 (step 6) followed by its hydroxylation with the release of a water molecule (step 7). The theoretical bases of the P450 catalytic cycle have been tested experimentally for decades (Poulos 2003). In order to engineer and elucidate P450 function, this cycle has to be taken into account.

4.3. Origin and evolution of eukaryotic P450s

Having been found in all kingdoms of life, including Archaea, the P450 superfamily seems to originate from an ancestral gene that existed more than three billion years ago (Danielson 2002). The cytochromes P450 family may previously have functioned as nitroreductases or endoperoxide isomerases upon the accumulation of significant levels of molecular oxygen. Then they may have evolved to protect early life forms from oxygen toxicity (Danielson 2002). Several gene duplications have given rise to one of the largest of multigene families. Interestingly, the CYP locus (from P450 coding region) is very close to the Homeobox genes (Hox, WNT and NK gene clusters), which underwent several duplication guaranteeing its present in all branches of the animal kingdom for example (Nelson, Goldstone et al. 2013). Homeobox genes are present in all eukaryotes and play critical roles in development in plants, fungi, and animals (Holland 2013). Hox and NK homeobox genes are present in plants and animals, while the WNT is specific from the metazoan group.

According to one theory, 1.5 billion years ago, the first of expansion of the P450 gene gave rise to the families of cytochromes P450 that are primarily involved in primary biosynthetic functions, for example fatty acids, cholesterol, etc. Another expansion of the gene family is believed to have occurred 900 million years ago, giving rise to several of the endogenous steroid-synthesizing cytochrome P450 lineages. These families later diverged into the major drug and carcinogen-metabolizing enzymes, which are nowadays the focus of intensive research in mammals (Danielson 2002).

Cytochromes P450 enzymes may have helped to keep membrane integrity of early eukaryotic cells during the first expansion and facilitated the conquest of land from plants later around 470 million years ago (Pires and Dolan 2012). Curiously, while the algae *Chlamydomonas* has around 50 P450 encoding genes, primitive land plants (bryophytes) have more than 70 genes (Hamberger and Bak 2013). Bryophytes had to adapt to high radiation levels, dehydration and to reproduce in a dry environment (Mizutani and Ohta 2010). This is how new P450 genes provided plants with an evolutionary advantage in the biosynthesis of several essential metabolites for survival on land (Nambara and Marion-Poll 2005). Another important diversification occurred in vascular plants, that can have up to 250 P450 encoding genes in some cases representing 1% of the entire genomes. This vast diversification of plant P450s did not arise separately from their animal counterpart, but as a consequence of a constant "warfare" for survival between this two groups. In plants, P450s are not only involved in biosynthetic functions such as hormones, lipid metabolism, etc. but also in the production of a wide range of chemical "weapons" against predators and the environment.

Animals followed plants in the conquest of land 440 million years ago according to the fossil record.

A dramatic expansion of several cytochrome P450 families involved in xenobiotic metabolism began about this time (Danielson 2002). Several factors are hypothesized to have triggered this early diversification of the P450 family in animals. Probably the most important is thought to be the land conquest by aquatic organisms resulting in their exposure to toxic plant allelochemicals into their diets (Danielson 2002). Examples of this can be found in the convergent evolution of the cyanogenic glucosides metabolic pathways of insects and plants (Jensen, Zagrobelny et al. 2011). In turn this process has pushed adaptation strategies of animals to avoid toxicity. For example in humans we only find 57 genes encoding for P450s enzymes, most of them involved in detoxification and drug metabolism. Humans adapted to xenobiotics with relatively few P450s, however, those display a very broad specificity for substrates. As plants evolved cytochromes P450 with the ability to synthesize toxic phytoalexins, many with antimicrobial,

antifungal, and insecticidal activities, to make themselves a less attractive food source, herbivorous animals responded by evolving new P450s capable of detoxifying these defensive plant allelochemicals (Danielson 2002). This is one of the possible reasons why P450s are so diversified in plants, particularly in vascular plants (Hamberger and Bak 2013).

Finally it is worth noticing that insects, especially mosquitoes, have the most P450 genes among the animal taxon. However, only a few of them are conserved across taxa for development metabolism. The vast majority are adapted to functions such as detoxification of insecticides. This is particularly important to humans because the number of people at risk for mosquito borne diseases (Nelson 2013). With this diversification landscape, enormous efforts for characterizing a significant number of plant P450s can be expected.

4.4. State-of-the-art for P450 expression and characterization

Because most eukaryotic P450s are membrane-bound, their characterization faces similar challenges as for other membrane proteins. Typically, P450s are expressed in different hosts, purified and characterized by their activity or absorption peak at 450nm. Often P450 engineering or the purification process yields an inactive absorption spectrum at 450nm. However, this does not necessarily imply that the enzyme is inactive, as we have experienced in some of the studies covered in this thesis. This generates a recurrent debate in the field over the usefulness of spectroscopy to estimate the activity or the amount of P450s. Attempts have been made to increase the high-throughput of this technique, however, it still represents a time-consuming operation to screen for a high number of variants. The main challenge in P450 characterization though, is expression levels of these enzymes in microbes (Chang, Eachus et al. 2007). Unfortunately there is not a clear and systematic approach to P450 expression. Based on these challenges, this Ph.D. thesis aims to use the lessons from the membrane protein expression field to propose a more standardized method for P450 expression and engineering. It is worth noticing that P450s are not integral membrane proteins, thus there is a need for developing new technologies or adapting the existing ones in order to accelerate the characterization and application of P450s. To date, two strategies are commonly used to boost expression of P450s in *E. coli* mentioned in the sections below.

4.4.1. Optimization of expression conditions

Expression conditions have been shown to impact both expression and activity of P450s. Factors such as temperature, growth time and media affect the yield of protein expression. Typically, bacteria carrying a recombinant cytochrome P450 plasmid are grown in Terrific

Broth media as it allows for high cell density growth (Zelasko, Palaria et al. 2013). Lower expression temperature has been reported to yield stable expression of P450s with minimal aggregation (Zelasko, Palaria et al. 2013). However, the low temperature usually leads to low expression and slower expression rates as well. Cytochrome P450 expression conditions are therefore typically performed at temperatures below 30°C for 24–48 h. Finally, media supplements such as the addition of d-aminolevulinic acid (ALA) during expression aims to assist heme biosynthesis in a many heterologous expression systems, and therefore is necessary sometimes for correct P450 holoenzyme formation (Zelasko, Palaria et al. 2013). ALA is an an amino acid derivative and the first common (?) intermediate in the heme, chlorophyll, cytochrome and vitamin B12 biosynthetic pathways. In bacteria, three enzymatic reactions catalyse the biosynthesis of ALA from glutamate (Zhang, Kang et al. 2015). Co-factor supply is a crucial point because the functional output of holoenzymes can only arise if the apoenzyme is correctly folded with its cofactor. It also important to consider co-factor supply in metabolic engineering applications because without the cofactor, enzymes would be inactive and the associated pathways would not lead to the desired product (Zhang, Kang et al. 2015).

4.4.2. Removal of the transmembrane segment

A way to avoid toxicity problems associated with the saturation of the translocation machinery in *E. coli* as well as unknown protein-lipid interactions, is the removal of the transmembrane segment. In section 2.3 we have described tools that predict transmembrane segment within a protein. Once the transmembrane segment is identified we can use DNA editing tools to truncate it from the encoding backbone. This rational strategy has been successfully applied for expression and characterization of several human, insect and plant P450s in *E. coli* (Budriang, Rongnoparut et al. 2011, Park, Lim et al. 2014, Gnanasekaran, Vavitsas et al. 2015). However, the success of the truncation strategy is case-dependent, and leads in some cases to inactive or minimally active enzyme forms (Mikkelsen, Hansen et al. 2000). Furthermore, the truncation point of the polypeptide sequence is not usually obvious, thus quite often several truncated points must be tested. This is a time-consuming process, as it requires cloning of the desired constructs and screening for activity. In the “omics” era such workflows still represent an important bottleneck. Finally, truncations have other side effects, such as breaking the protein association to the membrane (Gnanasekaran, Vavitsas et

al. 2015). When the purpose of a study is to structurally characterize an enzyme in its native environment, this may represent a drawback.

4.4.3. N-terminal modifications

Perhaps the most popular and widespread approach to achieve high expression of P450s in *E. coli* is not truncation, but addition of sequences or replacement of the transmembrane region. The popularity of this technique can be attributed mainly to Barnes and co-workers whom generated an AT-rich sequence, or Alanine substitution of the second codon, in the N-terminus of the bovine P450 17α-hydroxylase (Barnes, Arlotto et al. 1991). The resulting N-terminal sequence, commonly known as the Barnes sequence (MALLLAVFL), has been exploited in many other experiments and has become a golden standard of N-terminal modifications for P450 expression. The study from Barnes and co-workers was the first to specifically investigate the effect of DNA mutations of the N-terminus sequence on expression of functional P450s in *E. coli*. Several other P450s have been successfully expressed by N-terminus codon-optimization approaches (Zelasko, Palaria et al. 2013). Changes in the linker region between the SD and the ATG start codon have have a major effect on expression levels of the downstream ORF sequence (Mirzadeh, Martinez et al. 2015). It is speculated that thermodynamic parameter of this region may govern gene expression to some extent (Mirzadeh, Martinez et al. 2015). In the chapter 4 of this thesis we explore a technique developed for membrane protein expression, that relies on the randomization of this sequence.

In addition to these modifications, the fusion of signal peptides or short hydrophobic sequences represents another widely used strategy for improving P450 expression. This is to target heterologous protein specifically through the SRP-dependent pathway. As explained in section 3.2. the SRP mediate translocation of a nascent proteins with highly hydrophobic N-terminal sequences. The ultimate goal is to either target cargo protein to the inner membrane or to secrete it to the periplasmic space. Most popular variants of the strategy consist of fusion of short hydrophobic sequences, signal peptides or replacement of the transmembrane segment with the same, while the later have given the most promising results suggesting that other factors than the transmembrane segment itself may limit expression and activity levels (Gillam, Wunsch et al. 1997, Hull, Vij et al. 2000, Nielsen and Moller 2000, Leonard and Koffas 2007, Chang, Wang et al. 2010). Again, replacing the transmembrane segment with the Barnes sequence, have not been sufficient in all cases, and still requires a tedious process to engineer a functional P450 (Nielsen and Moller 2000). Some factors may explain the failure of this strategy, for example the thickness difference of

the *E. coli* membrane compared to the ER membrane that gives rise to the hydrophobic mismatch, disruption of local secondary structures or simply protein aggregation. The concept behind signal peptide fusions is to take advantage of the endogenous host translocation machinery in order to target the protein through the membrane as described in section 3. Although this strategy worked for characterization of the cytochromes P450s CYP3A4, CYP2A6, and CYP2E1, it has not been wide spread in the field (Pritchard, Ossetian et al. 1997). Furthermore these fusions have been also used for displaying P450 in the bacterial surface (Quehl, Hollender et al. 2016).

To sum up, there is no a systematic optimization workflow to tackle P450 expression. Several strategies have been proven successful, therefore in this thesis we outline some of the most popular strategies applied in the field and propose new tools to facilitate these optimization efforts using a GFP reporter platform.

4.5. Dhurrin biosynthesis as a model pathway

Dhurrin belongs to a group of natural compounds called cyanogenic glucosides, hydrogen cyanide-releasing chemicals that play important roles in defense against herbivore predation (Moller 2010). These defense chemicals are generally synthesized from amino acids from which they undergo several hydroxylation reactions catalyzed by P450s. Dhurrin is synthesized from tyrosine and released by plants to deter herbivores from feeding on leaves. The first two steps in the dhurrin pathway of *Sorghum bicolor* are catalyzed by the cytochromes P450 CYP79A1 and CYP71E1, respectively. The two intermediates generated by each enzyme, *p*-hydroxyphenylacetaldoxime and *p*-hydroxymandelonitrile, are known as oximes. The remaining pathway consist of the natural redox partner POR or CPR2b (Cytochrome P450 reductase) and a UDP-glucosil transferase (*Sb*UGT85B1) (Bak, Olsen et al. 2000). The two oximes generated through the dhurrin pathway can be separated by chromatography techniques such as thin-layer-chromatorgraphy (TLC) (Fig. 19). Therefore, dhurrin is an inexpensive and attractive proof-of-concept pathway to quickly assess the effect of P450 engineering strategies in a cell factory context (Nielsen, Ziersen et al. 2013).

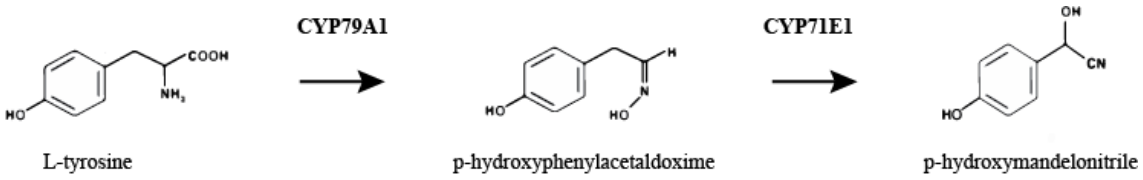


Fig. 19. Illustration of the P450-catalyzed steps of the dhurring pathway and the intermediates generated

CYP79A1 is well-studied and successfully expressed in several hosts, but the 3D protein structure has not been determined. Mutations of this enzyme have been studied in plants, and its interaction with the other enzymes of the pathway demonstrated with several experimental techniques. CYP79A1 can be regarded as a workhorse for metabolic engineering and structural studies. The first two chapters of this thesis outline technologies to de-bug expression of plant P450s in plants and how to tune expression for metabolic engineering applications, using, among others, the cytochromes of the dhurrin pathway. In the third chapter we show how these technologies can be deployed in order to elucidate previously uncharacterized structural elements of the CYP79A1.

4.6. Application of P450s in the biological production of natural medicines

In the previous sections we have described the general features and diversity of P450 enzymes. While human P450s are mostly involved in detoxification of compounds and biosynthesis of a narrow set of chemicals, plants as sessile organisms had to develop a vast arsenal of P450s to synthesise defence chemicals, hormones, pigments, etc. to adapt to its environment. It turns out that these defence chemicals, also have an important biological activity for humans. For example, antibiotics, antioxidants or inhibitors of cell growth that shows a high potential as drugs against many diseases and aging. Two examples of natural chemicals with medicinal potential, whose biosynthesis is dependent on the catalytic activities of P450s are here presented. The antimalarial drug artemisinin, which became the first plant chemical produced in cell factories, and taxol, a complex anti-cancer compound not yet possible to produce entirely in cell factories.

4.6.1. Artemisinin

The case of artemisinin production has been mentioned briefly in the section 2.1. but represents perhaps the most pragmatic and successful example of P450 enzymes used in cell factories. This initiative was motivated by instability of prices of the antimalarial drug in order to provide an alternative source of artemisinin that could stabilize the supply for patients in the developing countries. This project involved the metabolic engineering of microorganisms to produce the chemical precursor of artemisinin (artemisinic acid), which could be then converted to the final product by chemical synthesis, a product that is indistinguishable from plant-derived artemisinin. Interestingly much effort in this project was devoted to optimizing the oxidation of amorphadiene to artemisinic acid carried out naturally by the cytochrome P450 CYP71AV1 of *Artemisia annua*. Initially it was unclear whether this enzyme could provide a turn-over high enough for the commercial production

scale of the semi-synthetic artemisinin. However, the instability of one the reaction intermediates made impossible the use of alternative oxidizing enzymes such as the P450 BM3. Furthermore the production levels of amorphadiene in yeast were much lower than in *E. coli* (Paddon and Keasling 2014). Although studies in *E. coli* show promising levels of artemisinic acid (1g/L) thanks to modifications in the N-terminus of the CYP71AV1, it was considered a bad host for expression of eukaryotic P450s, thus further optimization steps were conducted in yeast (Chang, Eachus et al. 2007). This project showcases the need of further studies that can help improve the use of P450s in bacterial cell factories and their importance in the production of valuable chemicals.

4.6.2. Taxol

Taxol is a natural diterpenoid originally isolated from the stem bark of Pacific yew *Taxus brevifolia*. Under the commercial name Paclitaxel©, this chemical is of immense importance as an antineoplastic agent for the treatment of advanced, progressive and drug refractory ovarian cancer. The biosynthetic pathway is complex involving several oxidation steps catalysed by P450s, and yet, is not totally elucidated. The first studies on the hydroxylations of the taxane backbone were performed with cell-free extracts from *T. brevifolia* and *T. cuspidate* demonstrating that most of these reactions were catalyzed by microsomal P450s (Kaspera and Croteau 2006). The current production technology of taxol focuses on increase the productivity using plant cells and tissue cultures, and including optimisation of culture conditions, selection of high-producing cell lines, use of elicitors, and addition of precursors. The fact of containing multiple P450 enzymes complicates enormously the biosynthesis in microbes. The first step of taxol biosynthetic pathway is the cyclization of geranylgeranyl diphosphate (GGPP), which comes from the alternative pathway of the mevalonate. This cyclization leads to the formation of taxa-(4,5),(11,12)-diene, the first taxoid skeleton intermediate. This reaction is catalyzed by taxadiene synthase (TS), a terpene synthase (Exposito, Bonfill et al. 2009). The product of the reaction undergoes several hydroxylation rounds that deserve a detailed attention and are not fully covered in this thesis (Fig. 20) (Kaspera and Croteau 2006).

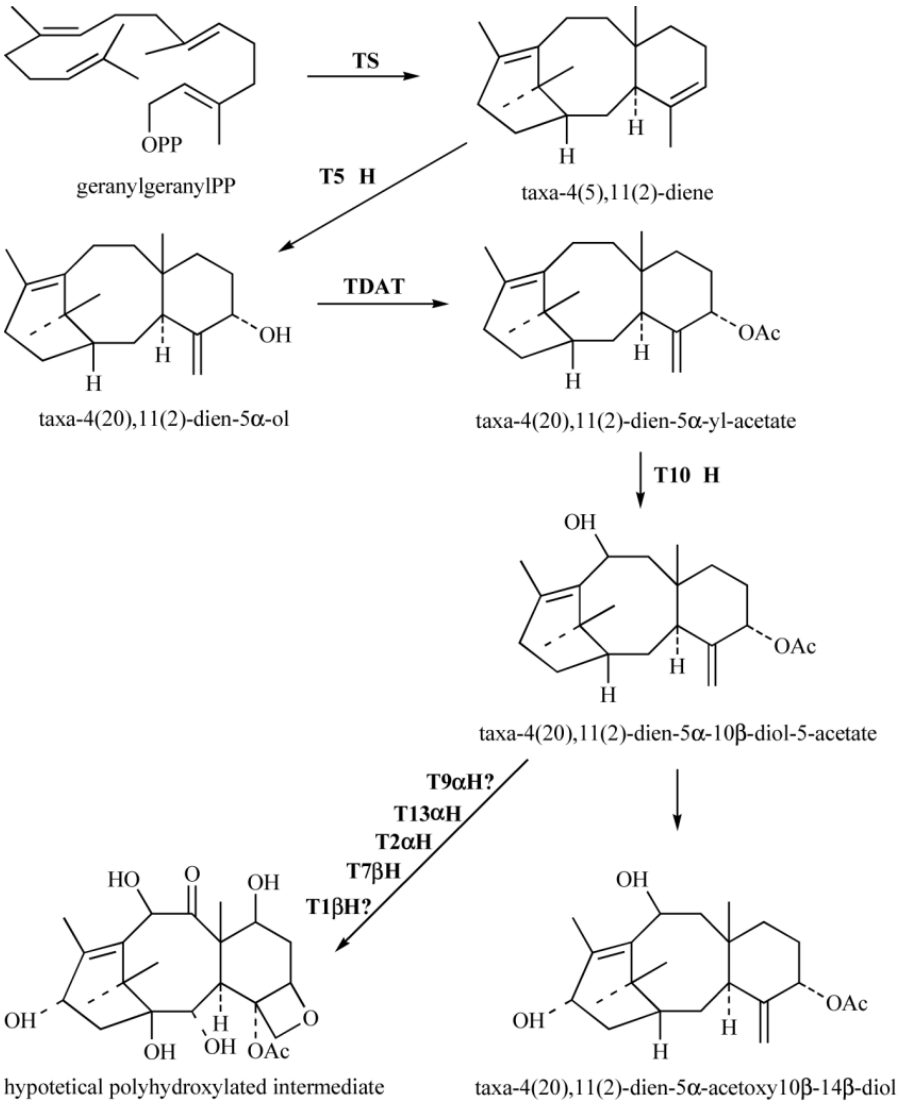


Fig. 20. First steps of the taxol biosynthetic pathway starting with GGPP (adapted) (Exposito, Bonfill et al. 2009).

Following the cyclization by the TS, the intermediate taxa-(4,5),(11,12)-diene is hydroxylated at the C5 position by a P450 taxadiene-5-hydroxylase (T5αH) to form taxa-4(20),11(12)-dien-5-ol. The next step in the biosynthetic pathway that leads to taxol is catalyzed by a specific taxadiene-5-ol-O-acetyl transferase (TDAT) that acylates taxa-4(20),11(12)-dien-5-ol at the C5 position to form taxa-4(20),11(12)-dien-5-yl-acetate. This taxoid is then hydroxylated by a P450-dependent monooxygenase, found in *T. cuspidate*, which catalyzes the hydroxylation of at the C10 position to yield taxa-4(20),11(12)-dien-5,10-diol 5-acetate (T10βH). Interestingly, other P450-dependent hydroxylases have been found that, using taxa-4(20),11(12)-dien-5-ol as a substrate, leads to the formation of taxa-4(20),11(12)-dien-5-13-diol suggesting that taxol biosynthesis is not a linear pathway and that there are branch points that can lead to other related taxoids. Although the order of the

hydroxylations in the taxol biosynthesis is not completely elucidated the most probable hydroxylation sequence based on *in vitro* cell cultures could be: C5, C10, C2, C9, C13 and finally C1. Lastly the final steps of the pathway involve the oxetane ring formation (essential for the anti-cancer activity of the taxol molecule), further acylation, and esterification with an α-phenylalanoyl-CoA molecule. The very last step of the pathway is thought to be catalyzed by an unknown P450s that hydroxylates the C2' position (Fig. 21).

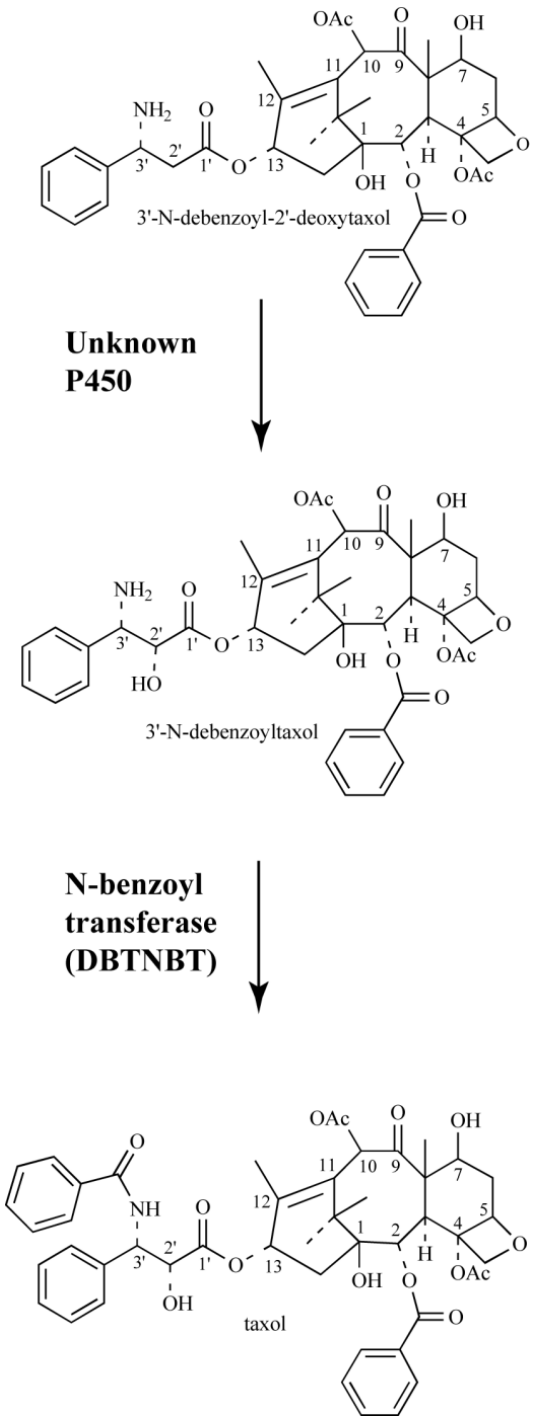


Fig. 21. Summary of the final steps of the taxol biosynthetic pathway (adapted) (Exposito, Bonfill et al. 2009).

This pathway illustrates the complexity of the different steps required for the biosynthesis of a powerful medicine not possible to obtain through classical organic synthesis. This highlights the importance of a robust P450 technology to develop cell factories capable of producing commercially relevant amounts of complex chemicals.

Definition of goals

Given the challenges of P450s expression and characterization, new technologies for screening a large number of conditions and engineered P450 variants in a quantitative or qualitative way are in great need. Screening variables should be ideally linked to expression levels, protein folding and enzymatic activity to assess the effect of the optimization conditions. Existing technologies in the membrane protein, structural biology, and metabolic engineering fields display advantages that could be useful to identify good P450 engineering strategies. The objectives of this PhD thesis are therefore:

- Develop a fluorescence-based high-throughput screening platform for expression of P450s in *Escherichia coli* strains
- Validate P450 activity of those highly expressed variants and conditions
- Systematically expand the toolbox of N-terminal tags that can boost P450 expression
- Build multigene metabolic pathways with N-terminally-engineered P450s
- Generate alternative methods for structural characterization of P450s using computational tools
- Combine computer-generated information with the high-throughput screening platform

In the first chapter we describe a systematic approach to tackle large P450 expression screenings and engineering using a GFP-based optimization platform from the membrane protein field. As we discover the usefulness of P450 expression associated to GFP fluorescence, in the second chapter we expand the number of N-terminal modifications of plant P450s. Here we also validate cellular localization and function of engineered P450 variants in a metabolic engineering context, using the dhurrin pathway as model. Furthermore, one of these N-terminal modifications are applied to a large P450 library in order to show the applicability of this strategy for other P450s. And finally the third chapter aims to find alternative structural methods to characterize P450s, combining homology modelling with our GFP-based platform to elucidate key amino acid residues of the CYP79A1 active site.

CHAPTER 1

De-bugging and maximizing plant cytochrome P450 production in Escherichia coli with a scalable GFP-based optimization scheme

De-bugging and maximizing plant cytochrome P450 production in *Escherichia coli* with a scalable GFP-based optimization scheme

Ulla Christensen[1], Dario V. Albacete[1], Tonja Wolff[1], Morten T. Nielsen[1], Scott James Harrison[1], Anders Holmgaard Hansen[1], Birger Lindberg Møller[2,3], Susanna Seppälä[1,4] and Morten H. H. Nørholm[1,4]

[1] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Denmark; [2] Plant Biochemistry Laboratory, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C, Copenhagen, Denmark; [3]Center for Synthetic Biology: bioSYNergy, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C, Copenhagen, Denmark.

ABSTRACT

Cytochromes P450 are attractive enzyme targets in biotechnology as they catalyze stereospecific C-hydroxylations of complex core skeletons at positions that typically are difficult to access by chemical synthesis. Membrane bound CYPs are involved in nearly all plant pathways leading to the formation of high value compounds. In the present study, we systematically optimize the heterologous expression of six different plant-derived CYP genes in *Escherichia coli*, using a high-throughput workflow based on C-terminal fusions to the green fluorescent protein. The six genes can be over-expressed in a variety of *E. coli* strains using standard growth media. Furthermore, sequences encoding a small synthetic peptide and a small membrane anchor markedly enhance the expression of all six genes. For one of the CYPs, the length of the linker region between the predicted N-terminal transmembrane segment and the soluble domain is varied, verifying the importance of this region for enzymatic activity. The described work describes how membrane bound CYPs are optimally produced in *E. coli* and thus adds this plant multi-membered key enzyme family to the toolbox for bacterial cell factory design.

1. Introduction

Escherichia coli is a model organism and a popular workhorse for protein production, with a comprehensive set of available genetic tools. An obvious extension to producing individual proteins for downstream applications is to produce a set of enzymes that catalyze an entire biosynthetic pathway. In our desire to advance from a petrochemical based society towards a bio-based society, it is critical that we deepen our understanding of how such heterologous biosynthetic pathways should be designed and maintained for the production of high value chemicals from renewable feedstock (Møller, 2014). Plant cytochrome P450 enzymes (CYPs) are central to this approach, as they are involved in nearly all pathways that lead to the formation of high value compounds such as terpenoids, alkaloids and phenylpropanoids (e.g. flavonoids,

isoflavonoids, chalcones, aurones and lignans). Many of these compounds are used as medicines, condiments, flavors and fragrances (Morant et al., 2003), with well known examples being the antimalarial diterpenoid artemisinin (Chang et al., 2007), the anti-cancer drug paclitaxel (Chau et al., 2004), the flavor compound vanillin (Gallage and Møller, 2015; Hansen et al., 2009) and the steviol glucoside based-sweeteners (Davies and Deroles, 2014). Typically, the membrane anchored CYPs catalyze stereospecific hydroxylations of complex core carbon skeletons, at positions that are difficult to access by *de novo* chemical synthesis. CYPs often show high substrate specificity and although members of this ancient multigene family are found in all domains of life, plants seem to be particularly enriched with often more than 250 CYPs in a single species (Nelson, Biochim Biophys Acta (2011) 1814: 14-18). Unfortunately, plant CYPs have the reputation of being notoriously difficult to produce in *E. coli* (Chang et al., 2007). Furthermore, in order to work efficiently, CYPs often need co-production of other membrane-associated proteins such as reductases (Eugster et al., 1992) or cytochromes b5 (Paddon et al., 2013; Zhang et al., 2007).

Previous approaches to improve plant CYP production in *E. coli* includes exchanging parts of the native amino-terminal region with parts of the amino-terminal region of a codon optimized bovine CYP17alpha; often referred to as the "Barnes sequence" (Bak et al., 1998a; 1997; Barnes et al., 1991; Leonard and Koffas, 2007). Similarly, bacterial expression of human CYPs have been aided by bacterial leader sequences such as the membrane translocation signals from *pelB* and *ompA* (Pritchard et al., 1997); for a recent review see (Zelasko et al., 2013)

Membrane protein production in *E. coli* has benefited greatly from the introduction of a streamlined green fluorescent protein-based pipeline for rapid and simple assessment of proper expression (Drew et al., 2006; 2005; 2001; Lee et al., 2014). Briefly, a carboxy-terminal GFP-fusion has many advantages: without interfering with membrane targeting and integration, it acts as an expression reporter as fluorescence indicates that the mRNA was translated in full. Importantly, if the upstream protein is mislocalized and subsequently aggregated, GFP fluorescence is completely quenched. Further, once folded, GFP remains fluorescent even during otherwise denaturing applications such as SDS-PAGE. In summary, the carboxy-terminal GFP provides a cheap, reliable and fast readout of the proper production of membrane proteins. Moreover, the fluorescent properties of such chimeric proteins can be used for screening the optimal solubilization conditions for downstream applications, such as structural studies (Kawate and Gouaux, 2006; Sonoda et al., 2011).

Currently, there is no established formula for the heterologous expression of CYPs in E. coli, bur rather literature is scattered with a wealth of different genetic manipulations and use of various strains and growth conditions. To enable efficient heterologous CYP production, there is an obvious need to systematically determine the proper expression conditions. Here, we demonstrate how functional expression of six different plant CYPs in E. coli is achieved by very simple means. The carboxy-terminal GFP-approach readily lends

itself to high throughput applications, and therefore the proposed setup is bound to be highly useful for large scale expression screens.

2. Materials and Methods

2.1 Bacterial strains

Escherichia coli strain NEB 5-alpha (New England BioLabs, Ipswich, USA) was used for cloning of PCR products and propagation of plasmids. The following *E. coli* strains were used for gene expression: Rosetta2(DE3) pLysS (Novagen, Merck Millipore, Germany), Bl21(DE3) pLysS (Promega, Madison, USA), C41(DE3) (Miroux and Walker, 1996), KRX (Promega, Madison, USA), MC4100(DE3)pLysS and MG1655(DE3)pLysS.

2.2 PCR and uracil excision

All DNA manipulations were performed using uracil excision technology as previously described (Nour-Eldin et al., 2006; Nørholm, 2010) and all oligonucleotides are listed in Table S1. PCR products were amplified in 50 µL reactions containing: 1 µL PfuX7 DNA polymerase, 0.2 mM dNTPs (Thermo Scientific, Waltham, USA), 1.5 mM $MgCl_2$, 0.5 µM forward oligonucleotide, 0.5 µM reverse oligonucleotide, Phusion® HF Reaction Buffer (New England BioLabs, Ipswich, USA) and 50 ng plasmid template. A touch-down PCR program was used for amplification: Step 1: 2 min 98°C; step 2:15 sec 98°C, 20 sec 65°C (-1°C per cycle), 45 sec per kb at 72°C (step 2 repeated 9 times until 55°C, then repeated 20 cycles at the annealing temperature 55°C); step 3: 5 min 72°C; step 4 hold at 10°C. PCR products were gel purified from 1% (w/V) agarose gel using NucleoSpin Gel and PCR Clean-up (Macherey-Nagel, Düren, Germany) and eluted in 10% TE buffer. Purified PCR products were incubated with 1 µL USER™ enzyme (New England BioLabs, Ipswich, USA) for 30 min at 37°C and subsequently mixed with linearized vector backbone in a molar ratio 3:1 and incubated at 18°C for 1 h. A Nanodrop spectrophotometer 2000 (Thermo Scientific, Waltham, USA) was used for estimation of PCR product and vector concentration. Approx. 5 µL of the assembled PCR product:vector solution was transformed into NEB 5-alpha chemically competent cells according to the manufacturer's protocol. Transformants were selected on Luria Bertoni (LB) agar plates supplemented with 50 µg/mL kanamycin, 25 µg/mL chloramphenicol or 10 µg/mL gentamycin. Colonies were screened for gene insert by colony PCR using OneTaq 2X Master mix (New England BioLabs, Ipswich, USA). Vectors were extracted and purified using QIAprep Spin Kit (Qiagen) and verified by DNA sequencing (Eurofins Genomics, Ebersberg, Germany).

2.3 DNA constructs

All DNA constructs were made with uracil excision as previously described (REF). Details of oligonucleotides, template DNA and references can be found in Tables S1 and S2 and below. Briefly, *AsiSI* restriction enzyme recognition sites, present in the kanamycin resistance cassettes, were deleted from all pET28a(+)-derived construct . *AsiSI* uracil excision compatible cloning cassettes were inserted between sequences encoding a TEV protease site, GFP and polyhistidine tag and either the 28-tag (Nørholm et al., 2013) (Fig. 1A), a Barnes-like N-terminal sequence MALLLAVF, SohB (residues 1-48, KDT39511) or YafU (residues 1-88, KEN61237) . Six genes encoding different plant CYPs (*Sorghum bicolor CYP51G1* (U74319*); Avena strigosa CYP51H10* (DQ680849); *Sorghum bicolor CYP71E1* (O48958); *Sorghum bicolor CYP79A1* (U32624); *Arabidopsis thaliana CYP79B2* (NM_120158) and *Picea sitchensis CYP720B4* (HM245403)) were PCR amplified with sequence specific oligonucleotides and cloned into the different pET28-*derived* vectors either treated with the restriction enzyme *AsiSI* (Thermo Scientific, Waltham, USA) and the nicking enzyme *Nb.BbvCI* (New England Biolabs, Ipswich, USA) or PCR amplified. The *sohB*- and *yafU-based E. coli* membrane anchors have their C-termini on the cytoplasmic side of the *E. coli* inner membrane (Daley et al., 2005) and therefore the plant *CYPs* cloned into the corresponding backbones had their predicted N-terminal anchors truncated to avoid translocation of the CYP into the periplasmic space. Truncations were designed using TMHMM v. 2.0 (Sonnhammer et al., 1998). A Strep-HRV3C tagged, codon optimized *Sorghum bicolor CPR2b* (Wadsäter et al., 2012) was cloned into pET28a(+)-*tev-gfp-his8* and *gfp* was subsequently deleted and the origin of replication was swapped with the corresponding parts from pSEVA63 (Silva-Rocha et al., 2013) by amplifying pSEVA63 with the oligo nucleotides 5´-ATCCGCTUTAATTAAAGGCATCAAATAAAAC-3' and 5´-ACTAGTCTUGGACTCCTGTTGATAGATC-3' and the pET28-based *cpr2b* construct with the oligo nucleotides 5´-AAGCGGAUCTACGAGTTGCATGATAAAGAAGACAGTC-3' and 5´-AAGACTAGUCAATCCGGATATAGTTCCTCCTTTCAG-3'. A truncated version of *E. coli lepB* was amplified from a previously described pGem1-Lep construct (Hessa et al., 2005) using the oligonucleotides 5´-ACTCGAGGAUGGCGAATATGTTTGCCCTGATTC-3' and 5´-ATCGCTGCUTCCAGGACCACCACTAGTCTCG-3' and combined with the pET28-based full length *CYP* constructs amplified with 5´-ATCCTCGAGUCTCCTTCTTAAAG-3' in combination with the gene specific forward oligo nucleotides.

2.4 Culture media and expression conditions

All strains were grown aerobically in liquid cultures. For plasmid propagation single colonies were grown over night in 2xYeast/Tryptone medium at 37°C 250 rpm. For single plasmid transformations, chemically competent cells were transformed with 20 ng plasmid according to the manufacturer's protocol. For co-transformation of plasmids for expression assays, electrocompetent cells were transformed with 20 ng of

plasmid DNA. Four different media were tested for expression, LB (1% tryptone; 0.5% yeast extract; 1% NaCl), Terrific Broth (TB) (1.2% tryptone; 2.4% yeast extract; 0.4% glycerol; 17 mM potassium phosphate (monobasic); 72 mM potassium phosphate (dibasic)); the defined rich medium PA-5052 (Studier, 2005) (50 mM Na_2HPO_4; 50 mM KH_2PO_4; 25 mM $(NH_4)_2SO_4$; 2 mM $MgSO_4$; 10 μM metals; 0.5% glycerol; 0.05% glucose; 0.2% alpha-lactose; 200 μg/mL amino acids E, D, K, R, H, A, P, G, T, S, Q, N, V, L, I, F, W and M). and minimal M9 (M9 salts; 2mM $MgSO_4$; 0.1mM $CaCl_2$; 0.2% glycerol; 0.2% glucose; 10 μM Fe^{3+}; 200 μg/mL amino acids E, D, K, R, H, A, P, G, T, S, Q, N, V, L, I, F, W and M). For media tests, over night cultures were prepared with each of the four media supplemented with 0.5% (w/V) D-glucose and appropriate antibiotics. The pre-cultures were subsequently inoculated into each of the corresponding media for expression. For other expression assays, over night cultures were made in 96 deep well plates, inoculating single colonies into 800 μL TB supplemented with 0.5% (w/V) D-glucose and appropriate antibiotics and grown at 30°C 250 rpm in Innova®44R incubator shaker system (5 cm orbital shaking) (New Brunswick Scientific, Eppendorf, USA). The optical density (OD) of the over night cultures was measured at Abs_{600nm} on Plate Reader SynergyMx (SMATLD) (BioTek, Winooski, USA). Over night cultures were inoculated into 5 mL fresh TB medium in 24 deep well plates to a final OD of 0.05. Cells were incubated for approx. 2 h at 37°C 250 rpm to an OD of 0.3-0.5. All strains were induced with a final conc. of 0.4 mM isopropyl ß-D-1-thiogalactopyranoside (IPTG, dioxane free, Thermo Scientific, Waltham, USA). Despite the autoinduction capacity of the PA-5052 medium, as described previously (Lee et al., 2014), IPTG was use for induction in this medium as well. The strain KRX was furthermore induced with a final conc. of 5 mM L-rhamnose (Sigma-Aldrich, St. Louis, USA). Cultures were subsequently incubated at 25°C 150 rpm for 3 h or 22 h.

2.5 Whole cell fluorescence measurements

Whole cell fluorescence was measured using 2 mL induced culture. The cells were harvested (2,500xg, 20 min) and resuspended in a total of 100 μL PBS buffer. The GFP fluorophore was allowed to form for 1 h at room temperature and then fluorescence was detected using excitation at 485nm and emission at 512nm with a window of +/- 9 nm, gain value 50, using plate reader SynergyMx SMATLD (BioTek, Winooski, USA).

2.6 SDS-PAGE

Whole cells were lysed for 1.5 h at room temperature in TRIS buffer (50 mM Tris HCl pH 7.5; 150 mM NaCl; 2 mM $MgCl_2$) with 250 U/mL Benzonase ® nuclease (Sigma-Aldrich, St. Louis, USA); 5 mg/mL lysozyme egg white powder (Amresco, US, Ohio) and the cOmplete ULTRA EDTA-free protease inhibitor cocktail (Roche, Basel, Switzerland). The cell lysate (OD_{600nm}0.1) was analyzed in parallel with PageRuler™ Prestained Protein Ladder 10-170K (Thermo Scientific, Waltham, USA) by standard SDS-page using Mini-PROTEAN® TGX™ 4-15% gels (Bio-Rad, Hercules, USA). In-gel fluorescence was detected on a G:BOX UV-table (Syngene, Cambridge, UK).

2.7 Enzyme activity measurements

For CYP79A1 enzymatic assays, cells were harvested by centrifugation 2,500xg 4°C for 10 min, washed once in 50 mM Potassium phosphate (KPi) buffer pH 7.5 and resuspended to 0.03 OD units per µL in 50 mM KPi buffer. The enzyme reaction was carried out in 30 µL volume consisting of 5 mM NADPH, 0.5 mM L-Tyrosine (Sigma-Aldrich, St. Louis, USA), 50 mM KPi buffer and 20 µL cell suspension. Cells were incubated at 30°C 400 rpm for 60 min. The product (*E*)-*p*-hydroxyphenylacetaldoxime (oxime) was extracted by adding 150 µL methanol followed by incubation at room temperature for 10 min. Cells debris was discarded twice by centrifugation (20,000xg, 10 min) and the supernatant was transferred into HPLC vials and stored at -20°C prior to analysis. The oxime was also extracted directly from cultures co-expressing *SbCYP79A1* and *SbCPR2b*. 100 µL culture was extracted with 100 µL methanol, as described above.

2.8 LCMS detection

The CYP79A1 product (*E*)-*p*-hydroxyphenylacetaldoxime (Mw 151.17) was detected and quantified by LC-MS. A chemically synthesized (*Z*)-*p*-hydroxyphenylacetaldoxime standard was kindly provided by Mohammed Saddik Motawie (University of Copenhagen, Department of Plant and Environmental Sciences). The two geometrical isomers (Z)-*p*-hydroxyphenylacetaldoxime and (E)-*p*-hydroxyphenylacetaldoxmie were detected in both the chemical oxime standard and the oxime sample product due to instability and chemical equilibrium. Thus for the purpose of total oxime quantification the chromatogram area of the both peaks were summed. LC-MS data was collected on a Bruker Evoq triple quadropole mass spectrometer equipped with an Advance UHPLC pumping system. Samples were held in the CTC PAL autosampler at a temperature of 10.0 °C during the analysis. Injections (2 µL) of the sample were made onto a Supellco Discovery HS F5-3 HPLC column (3 µm particle size, 2.1 mm i.d. and 150 mm long). The column was held at a temperature of 30.0 °C. The solvent system (flow rate: 1.0 ml/min) used was water with (A) 100mM ammonium formate and (B) acetonitrile using the following elution profile: 0.5 min 95% A/5% B, linear gradient to 50% A/50% B for 3.0 min, 1.5 min 50% A/50% B and re-equilibration for 2 min 95% A/5% B. The column eluent flowed directly into the heated ESI probe of the MS, which was held at 350°C and a voltage of 4500 V. SRM data was collected in centroid at unit mass resolution. Positive ion mode with Q1 set to monitor 152.70 *m/z*, Q3 set to monitor 136 *m/z* and Q2 set to a collision energy of 10.0eV, with an Argon pressure of 1.5 mTorr. The other MS settings were as follows, Sheath Gas Flow Rate of 40 units, Aux Gas Flow Rate of 40 units, Sweep Gas Flow Rate of 20 units, Ion Transfer Tube Temp was 350 °C.

3. Results

3.1 Full length expression of six plant CYPs in E. coli grown in standard media

Previously, it was shown that simple addition of a 84 nucleotide/28 amino acid tag (28-tag) to the amino-terminus greatly enhanced the expression of challenging membrane protein encoding genes in *E. coli* (Nørholm et al., 2013). Here, we used a similar design principle to sandwich plant CYPs between an amino-terminal 28-tag and a carboxyterminal GFP (Fig. 1A). Six CYP genes from four different plants were cloned into the pET28-*28tag-tev-gfp-his8* plasmid: *Sorghum bicolor CYP51G1* (Bak et al., 1997),*CYP71E1* (Bak et al., 1998a) and *CYP79A1* (Koch et al., 1995), *Avena strigosa CYP51H10* (Qi et al., 2006), *Arabidopsis thaliana CYP79B2* (Bak et al., 1998b)*;* and *Picea sitchensis CYP720B4* (Hamberger et al., 2011) and transformed into *E. coli* KRX. Following three hours induction in four different standard growth media (LB, TB, M9 supplemented with $Fe^{3+}$and the synthetic rich medium PA-5052); expression and protein integrity was analyzed by measuring whole cell- and in-gel fluorescence. Compared to an empty vector control, all six constructs exhibited significantly elevated fluorescence levels in all growth media except M9 (Fig. 1B). Moreover, analysis of the samples by in-gel fluorescence indicated that the majority of the constructs were expressed as full-length proteins (Fig. 1C), except for *CYP51H10* for which a significant proportion seemed to be truncated or partly degraded. Addition of 5-aminolevulinic acid, suggested to aid the functional production of heme-containing proteins (Sudhamsu et al., 2010) did not have any apparent effect on expression levels (data not shown). Based on the highest expression level and cell density, we chose TB medium for further expression optimization.

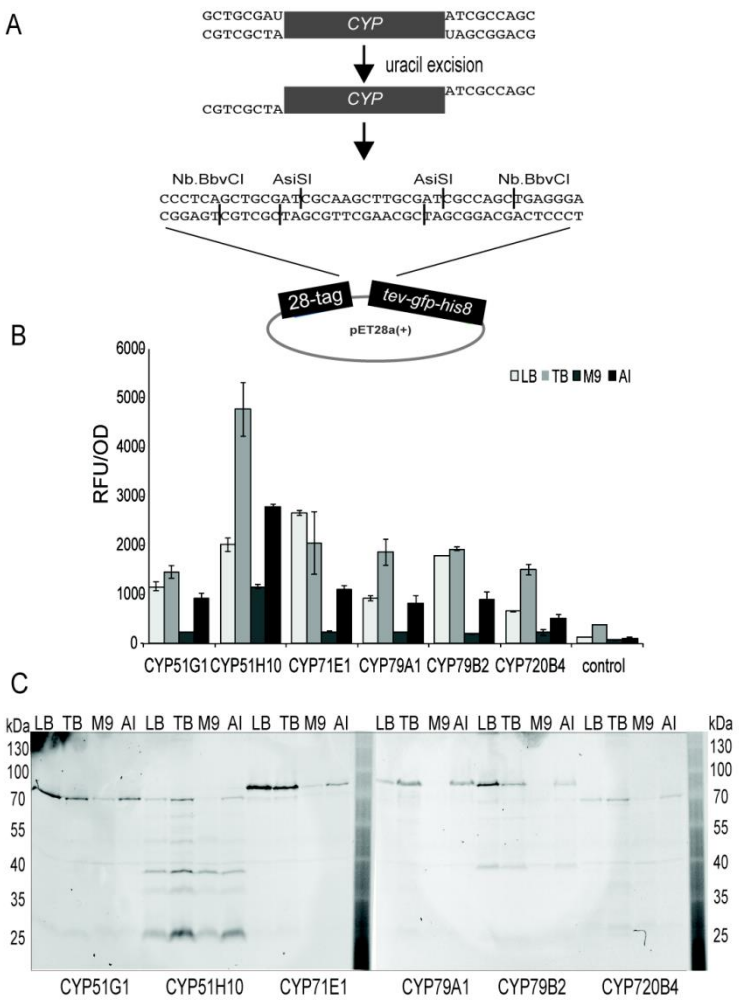


Fig. 1. Production of six different CYP enzymes in standard growth media. (A) Schematic representation of the cloning procedure used. Genes encoding six different CYPs were PCR amplified with specific uracil-containing tailed oligonucleotides and assembled with a compatible *AsisI/Nb. NbvcI* treated vector. (B) Effect of media on production of CYP51G1, CYP51H10, CYP71E1, CYP79A1, CYP79B2 and CYP720B4 in the KRX strain. Four media types were tested: Luria Bertoni (LB), Terrific Broth (TB), minimal media M9 and autoinducing media (AI). As cell cultures reached the exponential growth phase, cultures were induced for 3 h. Cell cultures were subsequently harvested by centrifugation and resuspended in PBS buffer for whole cell fluorescence detection (relative fluorescence unit (RFU) per optical density (OD)). Error bars indicate standard error of the mean (n=3). Empty vector (control) was included as reference. (C) The level of full-length GFP-fused protein monitored from cell lysate by in-gel fluorescence detection.

3.2 E. coli strains KRX and Rosetta excel in plant CYP expression

Next we compared expression levels of the six different CYP constructs in the six different *E. coli* strains MC4100(DE3) pLysS, MG1655(DE3) pLysS, KRX, BL21(DE3) pLysS, Rosetta2(DE3) pLysS , and C41(DE3). In all strains except for KRX and C41, expression from the T7 promoter is catalyzed by the T7-phage derived RNA polymerase controlled by the PlacUV5 IPTG-inducible promoter. In three of the strains, potential toxic effects of over-expression is counteracted by expression of the natural T7 DNA polymerase inhibitor T7 lysozyme from the pLysS or the pRARE plasmid (Studier, 1991), whereas a similar effect is obtained in C41 by mutations in PlacUV5 that make the promoter less strong (Wagner et al., 2008). In the commercial strain KRX expression of T7 RNA polymerase is tightly controlled by a rhamnose inducible promoter. The six pET28-28tag-*CYP-tev-gfp-his8* constructs were transformed into the six different strains and expression was monitored by whole cell- (Fig. 2A) and in-gel- (Fig. 2B) fluorescence after 3 h induction. As judged from whole cell fluorescence and presence of full-length protein on SDS-gels, strains KRX and Rosetta2(DE3) pLysS clearly outperformed the other four strains (Fig. 2A and 2B).

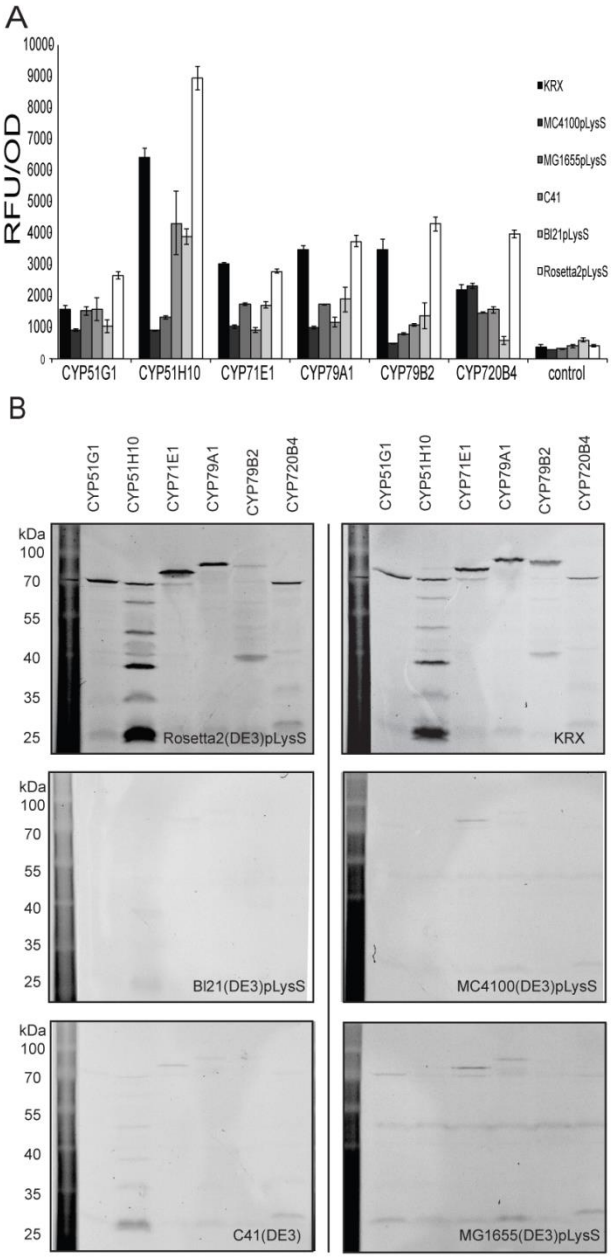


Fig. 2. Strain-dependent production of CYP enzymes. (A) Three K-strains (KRX, MC4100pLysS, MG1655pLysS) and three B-strains (C41, Bl21pLysS, Rosetta2pLysS) were transformed with *28-tag:CYP:tev-gfp-his8* expression constructs. Cells were grown in TB media and induced for 3 h, as cell cultures reached the exponential growing phase. Overall expression was monitored by whole cell fluorescence detection (relative fluorescence unit (RFU) per optical density (OD)). Error bars indicate standard error of the mean (n=3). Empty expression vector was included as control. (B) The level of full-length GFP-fused protein was monitored from cell lysate by in-gel fluorescence detection in the three B-strains (left panel) and the three K-strains (right panel).

3.3 A transmembrane domain encoded by E. coli sohB normalizes expression of plant CYP genes to high levels

Most eukaryotic CYPs utilize N-terminal hydrophobic peptides to localize to the endoplasmic reticulum, and sequence modifications in this region have previously proven essential for heterologous expression. We hypothesized that it would be beneficial to exchange the plant ER/membrane localization signals from the CYPs with membrane anchors that are derived from the heterologous expression host and that have previously been shown to express at high levels from the T7 promoter. To find suitable candidates, we mined a previously published library of membrane protein encoding genes expressed in *E. coli* (Daley et al., 2005). We chose three of the most highly expressed genes (Fig. 1S.) that encoded membrane proteins with different topologies: *sohB*, encoding a single-pass membrane protein with the C-terminal end localized inside the bacterial cytoplasm (C_{in}), *yafU*, encoding two membrane spanning regions with a C_{in} topology; and *lepB*, encoding two membrane spanning regions, but with the opposite (N_{out}) topology (Fig. 3A). To mimic the overall topology of the microsomal plant CYPs and ensure localization of the CYP catalytic domain to the cytoplasm, we replaced the hydrophobic sequences of the CYPs with the transmembrane parts of SohB and YafU; whereas the full length CYP was fused to the periplasmic carboxy-terminus of LepB. For benchmarking purposes, we added a Barnes-like MALLAVF-encoding sequence to the six *CYP* genes and comparisons were made to unmodified plant CYP sequences. The native controls, *28*-tag-, *MALLAVF*-, *sohB*-, *yafU*- and *lepB*-tagged *CYP* constructs were transformed into KRX and Rosetta2(DE3) pLysS and expression was monitored by whole cell (Fig. S2) and in-gel fluorescence (Fig. 3B). None of the native N-terminally unmodified sequences expressed to any significant levels. In contrast, both the 28-tag and the SohB-domain normalized the expression of the *CYPs* to high levels and in particular the combination of the KRX strain with the SohB-tagged CYPs consistently gave high levels of full-length protein. Furthermore, compared to the 28-tag constructs, a higher proportion of SohB-tagged CYP79A appeared as full-length, both 3 h and 22 h-post-induction (Fig. S3)

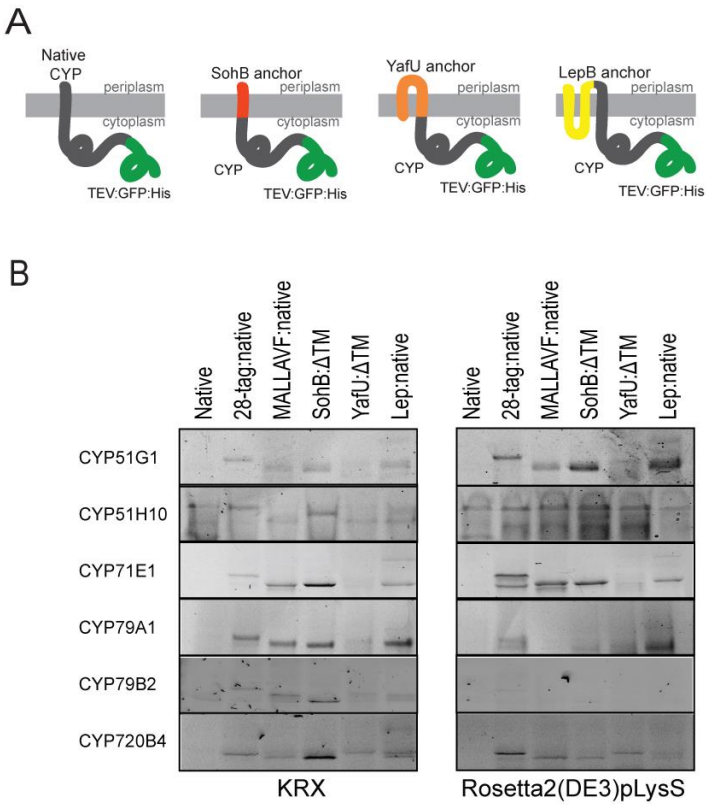


Fig. 3. Effect of different N-terminal peptides and small bacterial membrane anchors on the expression of six different *CYPs*. (A) Illustration of the hypothetical membrane topology of native CYPs and recombinant bacterial anchor constructs: (from left) reference CYP with native N-terminal sequence, SohB membrane domain (1TM, highlighted in red) fused with a truncated CYP, YafU anchor (2TM, C_{in}, highlighted in orange) fused with a truncated CYP, and LepB anchor (2TM, C_{out}, highlighted in yellow) fused to a native CYP. All constructs have TEV:GFP:His8 fusions at their C-terminal. (B) Six CYPs were tested with different N-terminal modifications: (from left) native sequence as reference, 28-tag fused to native CYP sequence, Barnes–like MALLAVF peptide fused to native CYP sequence, SohB membrane domain fused to a truncated CYP sequence, YafU membrane domain fused to a truncated CYP sequence and the LepB membrane domain fused to a native sequence. Constructs were expressed in two strains (KRX and Rosetta2(DE3)pLysS). Cell cultures were induced for 3 h before harvest. Cell lysates were analysed by in-gel fluorescence to determine the level of full-length protein present.

3.4 Activity is preserved in all engineered versions of CYP79A1 except for those truncated in a proline rich region that precedes the soluble catalytic domain

To obtain detectable expression levels of CYPs, we substantially re-engineered plant CYP proteins with truncations and/or heterologous protein domains such as the five different amino-terminal peptides and the large carboxy-terminal GFP domain described above. To test that a CYP is able to retain activity with these major modifications, we functionally assayed all the modified versions of CYP79A1 both *in vivo* and *in vitro* while monitoring expression levels by fluorescence. Moreover, we checked the effect of truncating CYP79A1 at the residue Y36, likely placed at the interface between the membrane and the aqueous face; at the residue P69, placed at a proline-rich, proposed hinge region observed in many P450s (Chen et al., 1998; Leonard and Koffas, 2007; Williams et al., 2000); and at two positions in between these two extremes (P44 and P60, Fig. 4A). In the *in vitro* assay, *CYP79A1* and the reductase *Sorghum bicolor CPR2b* were co-expressed in the KRX strain and the activity was measured on harvested cells by adding NADPH and the substrate Tyrosine. Formation of the product (*E*)-*p*-hydroxyphenylacetaldoxime (Fig. 4B) was followed by separation using high performance liquid chromatography and mass detection (LCMS) Fig.S4 and S5A). In parallel, we performed a quantitative *in vivo* assay by co-expressing *CYP79A1* with *SbCPR2b* and monitoring oxime formation with LCMS detection (Fig. 4C). Whole cell fluorescence confirmed expression of all constructs except the construct with the native, unmodified *CYP79A1* sequence, as also described above (Fig. S5B), and most of the expressed constructs were catalytically active except for those with the most heavily truncated CYP79A1 sequence at position P69. Cells expressing the *YafU* and *LepB* fusions were less active than those expressing the *28*- and the *SohB*-fusions (Fig. 4B). In order to see if the distance to the membrane is important, we increased the linker between the membrane anchor and the soluble domain by duplicating the sequence Y36-P44 in the SohB-CYP79A1 fusion construct. The artificially increased size of the linker resulted in a seemingly small increase in enzyme activity (Fig. 4B). Finally, particularly because the CYP needs to functionally interact with its reductase partner, we tested if the GFP fusion had a negative impact on the proper interactions, but we only observed minor improvements in the activity when GFP was removed (Fig. S6).

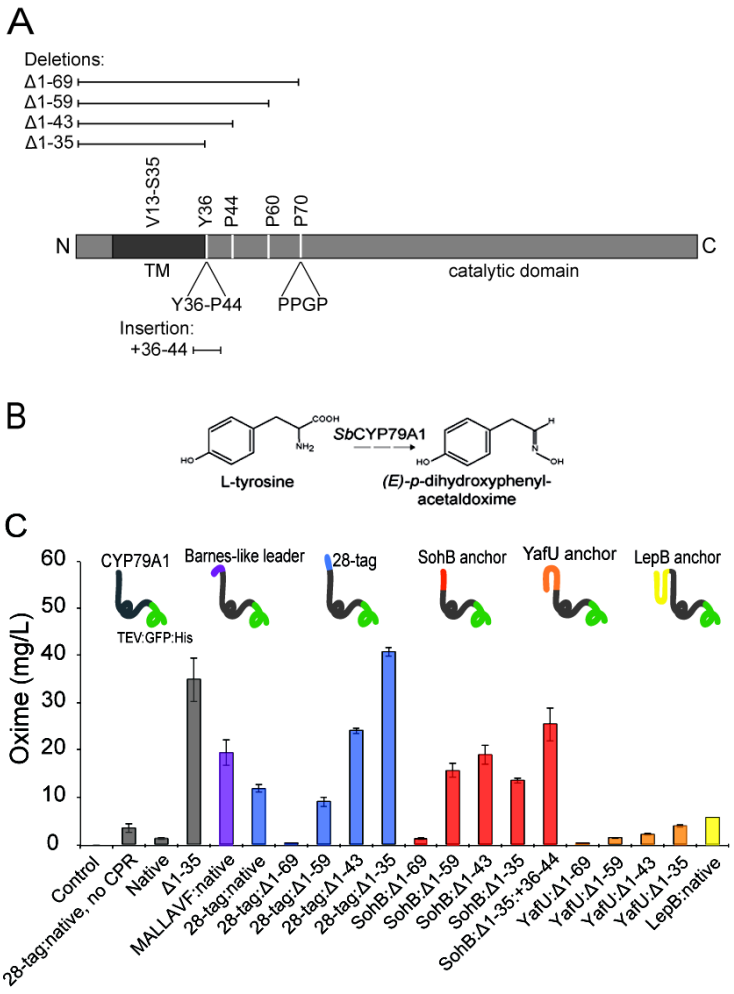


Fig. 4. Functional assaying of engineered variants of CYP79A1. (A) Schematic overview of CYP79A1 modified expression constructs used in this study. Four truncated expression constructs were designed to remove various parts of the N-terminal sequence: Δ1-35: at the end of the predicted transmembrane segment between residues 1 and 35, and at three proline residues at positions 44, 60 and 70 (Δ1-43, Δ1-59, Δ1-69) preceding the catalytic domain. In addition, an artificially extended linker region was engineered by inserting a repeat of the amino acids Y36-P44. (B) CYP79A1-catalyzed conversion of L-Tyrosine to *(E)-p*-hydroxyphenylacetaldoxime. (C) Differently truncated and engineered CYP79A1 constructs were co-produced with the compatible reductase electron donor CPR2b in the KRX strain. Cells were induced for 22 h, before oxime was extracted from the culture. Empty vector control and *28-tag:native CYP79A1* expressed without *CPR* were included as references. Error bars indicate standard error of the mean (n=3).

4. Discussion

The multitude of available genetic tools, growth media, and genotypes for the model organism *E. coli* is a blessing, but can also be a curse. Optimization of the physical parameters for heterologous gene expression is a multifactorial problem, and the details - from transcription to translation and post-translational processing - can be endlessly tinkered with. To this end, a fast and inexpensive assay to optimize expression conditions in a high-throughput fashion is of great importance, especially for the design of cell factories that are based on expression of multiple challenging enzymes like CYPs. Membrane bound CYPs are involved in nearly all plant pathways leading to the formation of high value compounds. GFP has been demonstrated to be a useful reporter of proper expression and protein folding for soluble proteins and membrane proteins alike (Drew et al., 2006; 2005; 2001; Lee et al., 2014; Waldo et al., 1999).

Here, we have explored the use of GFP fusions to report on the functional production of a commercially attractive group of single spanning membrane proteins, the multi-gene encoded cytochrome P450 enzyme family (Morant et al., 2003). We discovered that the two expression strains KRX and Rosetta2(DE3) pLysS consistently outperforms the other tested expression strains with respect to the amount of functionally active P450 enzyme produced under short term induction conditions. KRX is a K strain and Rosetta2 is a B strain, suggesting that there is no obvious restriction in the use of K vs. B strains for *CYP* expression. The Rosetta2 strain provides additional copies of tRNAs that are present in low concentration in *E. coli* strains, whereas KRX does not, suggesting that this complementation is not the major causative effect on *CYP* expression in Rosetta2. Rather, a likely similarity between the two strains is the carefully balanced expression of the T7 RNA polymerase – in KRX by controlling expression directly from the Prha promoter and in Rosetta2 indirectly by expression of *lysS* from the pRARE plasmid. Several of the other strains express *lysS* from the pLysS plasmid, but these do not perform as well, and we speculate that the difference in lysS carrying plasmid may change the production of lysozyme in a way that affects *CYP* expression. Our comparison of growth media suggests that there is no "magic" medium for *CYP* expression, but that the minimal medium M9 without additional supplements is suboptimal.

Another finding is that expression levels can be normalized to high levels using small peptides such as the 28-tag or membrane-spanning domains like SohB. Similar effects have been observed before with tags such as polyhistidine, the maltose binding protein (MBP) and the and small ubiquitin related modifier (SUMO, for a recent review see (Costa et al., 2014)). The frequently observed positive effect of adding sequences to the 5´-end is possibly due to an incompability between the native plant 5´sequences with high-level expression, as previously reported for membrane protein encoding genes like *araH* and *narK* (Nørholm et al., 2013). An alternative solution that previously has proven successful is to introduce synonymous changes in the first couple of codons downstream from the start codon. Indeed, we have successfully expressed the native

CYP79A1 and *CYP71E1* genes by introducing a few codons changes in the pET expression system (data not shown). We hypothesized that the inherent membrane targeting of e.g. SohB would have a positive impact of expression *and* activity of an ER-derived protein targeted to the *E. coli* inner membrane, but found no significant enhancement in activity when comparing SohB with e.g. 28-tagged CYPs. However, we also did not observe any major changes in activity (Fig.4b) or membrane localization of CYP79A1 even when completely removing all N-terminal (predicted) transmembrane segments (data not shown). Membrane association in the absence of N-terminal hydrophobic sequence has been observed before (Doray et al., 2001), suggesting that some CYPs are sufficiently hydrophobic to localize to the membrane in the absence of an N-terminal localization signal (Jensen et al., 2011). This is also supported by molecular dynamics studies suggesting interactions between helices located far downstream from the N-terminus and biological membranes (Denisov et al., 2012). The identification of SohB as a generic tool to enhance expression, while being native to the *E. coli* membrane insertional machinery, may become useful when engineering higher order membrane assemblies such as the suggested multi-CYP metabolons (Laursen et al., 2015; Moller, 2010). Also, even though the *lepB* fusions expressed to a lower level than those based on e.g. SohB, the chimeric proteins were active and may become useful because they maintain the native plant membrane anchor and (presumably) topology. This is particularly relevant because our results clearly demonstrate that care should be taken when truncating CYPs, as exemplified with the inactive P69-truncated CYP79A1.

5. Conclusion

The presented work demonstrates the usefulness of the GFP-reporter approach for de-bugging and maximizing expression of a group of related enzymes, and suggests that there are no inherent limitations in using different standard *E. coli* strains and expression conditions for exploiting CYPs for biotechnological applications. Further, the peptides and membrane domains used in this study add new biobricks to a toolbox for engineering pathways of higher complexity such as the taxol biosynthetic pathway consisting of eight steps catalyzed by CYPs (Chau et al., 2004).

Acknowledgements

We thank Victor de Lorenzo and the members of his laboratory for generously providing the pSEVA collection. We thank Tomas Laursen, Peter Naur, Søren Bak, Björn Hamberger, Johan Andersen-Ranberg and Britta Hamberger for advice on CYPs and reductases. We thank David Drew, Daniel Daley and Jan Willem de Gier for discussions on *E. coli* gene expression and the GFP-based expression platform. SS is the recipient of VILLUM Foundation's Young Investigator Programme grant VKR023128. This work was supported by the Novo Nordisk Foundation, from the VILLUM research center of excellence "Plant Plasticity", from the UCPH Excellence Program for Interdisciplinary Research to Center of Synthetic

Biology ”bioSYNergy” and by an European Research Council Advanced Grant Project No. 323034: LightdrivenP450s.

References

Bak, S., Kahn, R.A., Nielsen, H.L., Moller, B.L., Halkier, B.A., 1998a. Cloning of three A-type cytochromes P450, CYP71E1, CYP98, and CYP99 from Sorghum bicolor (L.) Moench by a PCR approach and identification by expression in Escherichia coli of CYP71E1 as a multifunctional cytochrome P450 in the biosynthesis of the cyanogenic glucoside dhurrin. Plant Mol. Biol. 36, 393–405.

Bak, S., Kahn, R.A., Olsen, C.E., Halkier, B.A., 1997. Cloning and expression in Escherichia coli of the obtusifoliol 14 alpha-demethylase of Sorghum bicolor (L.) Moench, a cytochrome P450 orthologous to the sterol 14 alpha-demethylases (CYP51) from fungi and mammals. Plant J. 11, 191–201.

Bak, S., Nielsen, H.L., Halkier, B.A., 1998b. The presence of CYP79 homologues in glucosinolate-producing plants shows evolutionary conservation of the enzymes in the conversion of amino acid to aldoxime in the biosynthesis of cyanogenic glucosides and glucosinolates. Plant Mol. Biol. 38, 725–734.

Barnes, H.J., Arlotto, M.P., Waterman, M.R., 1991. Expression and enzymatic activity of recombinant cytochrome P450 17 alpha-hydroxylase in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 88, 5597–5601.

Chang, M.C.Y., Eachus, R.A., Trieu, W., Ro, D.-K., Keasling, J.D., 2007. Engineering Escherichia coli for production of functionalized terpenoids using plant P450s. Nat. Chem. Biol. 3, 274–277. doi:10.1038/nchembio875

Chau, M., Jennewein, S., Walker, K., Croteau, R., 2004. Taxol biosynthesis: Molecular cloning and of a cytochrome p450 characterization taxoid 7 beta-hydroxylase. Chemistry & Biology 11, 663–672. doi:10.1016/j.chembiol.2004.02.025

Chen, C.D., Doray, B., Kemper, B., 1998. A conserved proline-rich sequence between the N-terminal signal-anchor and catalytic domains is required for assembly of functional cytochrome P450 2C2. Arch. Biochem. Biophys. 350, 233–238. doi:10.1006/abbi.1997.0524

Costa, S., Almeida, A., Castro, A., Domingues, L., 2014. Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system. Front Microbiol 5, 63. doi:10.3389/fmicb.2014.00063

Daley, D.O., Rapp, M., Granseth, E., Melen, K., Drew, D., Heijne, von, G., 2005. Global topology analysis of the Escherichia coli inner membrane proteome. Science 308, 1321–1323. doi:10.1126/science.1109730

Davies, K.M., Deroles, S.C., 2014. Prospects for the use of plant cell cultures in food biotechnology. Current Opinion in Biotechnology 26, 133–140. doi:10.1016/j.copbio.2013.12.010

Denisov, I.G., Shih, A.Y., Sligar, S.G., 2012. Structural differences between soluble and membrane bound cytochrome P450s. J. Inorg. Biochem. 108, 150–158. doi:10.1016/j.jinorgbio.2011.11.026

Doray, B., Chen, C.D., Kemper, B., 2001. N-terminal deletions and His-tag fusions dramatically affect expression of cytochrome p450 2C2 in bacteria. Arch. Biochem. Biophys. 393, 143–153. doi:10.1006/abbi.2001.2473

Drew, D., Lerch, M., Kunji, E., Slotboom, D.J., de Gier, J.W., 2006. Optimization of membrane protein overexpression and purification using GFP fusions. Nat Methods 3, 303–313. doi:10.1038/nmeth0406-303

Drew, D., Slotboom, D.-J., Friso, G., Reda, T., Genevaux, P., Rapp, M., Meindl-Beinker, N.M., Lambert, W., Lerch, M., Daley, D.O., Van Wijk, K.-J., Hirst, J., Kunji, E., De Gier, J.-W., 2005. A scalable, GFP-based pipeline for membrane protein overexpression screening and purification. Protein Science 14, 2011–2017. doi:10.1110/ps.051466205

Drew, D.E., Heijne, von, G., Nordlund, P., de Gier, J.W., 2001. Green fluorescent protein as an indicator to monitor membrane protein overexpression in Escherichia coli. FEBS Lett. 507, 220–224.

Eugster, H.P., Bärtsch, S., Würgler, F.E., Sengstag, C., 1992. Functional co-expression of human oxidoreductase and cytochrome P450 1A1 in Saccharomyces cerevisiae results in increased EROD activity. Biochem. Biophys. Res. Commun. 185, 641–647.

Gallage, N.J., Møller, B.L., 2015. Vanillin-Bioconversion and Bioengineering of the Most Popular Plant Flavor and Its De Novo Biosynthesis in the Vanilla Orchid. Mol Plant 8, 40–57. doi:10.1016/j.molp.2014.11.008

Hamberger, B., Ohnishi, T., Hamberger, B., Séguin, A., Bohlmann, J., 2011. Evolution of diterpene metabolism: Sitka spruce CYP720B4 catalyzes multiple oxidations in resin acid biosynthesis of conifer defense against insects. Plant Physiol 157, 1677–1695. doi:10.1104/pp.111.185843

Hansen, E.H., Møller, B.L., Kock, G.R., Bünner, C.M., Kristensen, C., Jensen, O.R., Okkels, F.T., Olsen, C.E., Motawia, M.S., Hansen, J., 2009. De novo biosynthesis of vanillin in fission yeast (Schizosaccharomyces pombe) and baker's yeast (Saccharomyces cerevisiae). Appl Environ Microbiol 75, 2765–2774. doi:10.1128/AEM.02681-08

Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., Heijne, von, G., 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature 433, 377–381. doi:10.1038/nature03216

Jensen, K., Osmani, S.A., Hamann, T., Naur, P., Møller, B.L., 2011. Homology modeling of the three membrane proteins of the dhurrin metabolon: catalytic sites, membrane surface association and protein-protein interactions. Phytochemistry 72, 2113–2123. doi:10.1016/j.phytochem.2011.05.001

Kawate, T., Gouaux, E., 2006. Fluorescence-Detection Size-Exclusion Chromatography for Precrystallization Screening of Integral Membrane Proteins. Structure 14, 673–681.

doi:10.1016/j.str.2006.01.013

Koch, B.M., Sibbesen, O., Halkier, B.A., Svendsen, I., Moller, B.L., 1995. The primary sequence of cytochrome P450tyr, the multifunctional N-hydroxylase catalyzing the conversion of L-tyrosine to p-hydroxyphenylacetaldehyde oxime in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench. Arch. Biochem. Biophys. 323, 177–186.

Laursen, T., Møller, B.L., Bassard, J.-E., 2015. Plasticity of specialized metabolism as mediated by dynamic metabolons. Trends Plant Sci. 20, 20–32. doi:10.1016/j.tplants.2014.11.002

Lee, C., Kang, H.J., Hjelm, A., Qureshi, A.A., Nji, E., Choudhury, H., Beis, K., De Gier, J.-W., Drew, D., 2014. MemStar: A one-shot Escherichia coli-based approach for high-level bacterial membrane protein production. FEBS Lett. 588, 3761–3769. doi:10.1016/j.febslet.2014.08.025

Leonard, E., Koffas, M.A.G., 2007. Engineering of artificial plant cytochrome p450 enzymes for synthesis of isoflavones by Escherichia coli. Appl Environ Microbiol 73, 7246–7251. doi:10.1128/AEM.01411-07

Miroux, B., Walker, J.E., 1996. Over-production of proteins in Escherichia coli: Mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. J. Mol. Biol. 260, 289–298. doi:10.1006/jmbi.1996.0399

Moller, B.L., 2010. Dynamic metabolons. Science.

Morant, M., Bak, S., Møller, B.L., Werck-Reichhart, D., 2003. Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. Current Opinion in Biotechnology 14, 151–162.

Møller, B.L., 2014. Synthetic Biology. Royal Society of Chemistry.

Nour-Eldin, H.H., Hansen, B.G., Nørholm, M.H.H., Jensen, J.K., Halkier, B.A., 2006. Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. Nucleic Acids Res. 34, e122–e122. doi:10.1093/nar/gkl635

Nørholm, M.H.H., 2010. A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. BMC Biotechnol. 10, 21. doi:10.1186/1472-6750-10-21

Nørholm, M.H.H., Toddo, S., Virkki, M.T.I., Light, S., Heijne, von, G., Daley, D.O., 2013. Improved production of membrane proteins in Escherichia coli by selective codon substitutions. FEBS Lett. 587, 2352–2358. doi:10.1016/j.febslet.2013.05.063

Paddon, C.J., Westfall, P.J., Pitera, D.J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M.D., Tai, A., Main, A., Eng, D., Polichuk, D.R., Teoh, K.H., Reed, D.W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., Diola, D., Dorin, G., Ellens, K.W., Fickes, S., Galazzo, J., Gaucher, S.P., Geistlinger, T., Henry, R., Hepp, M., Horning, T., Iqbal, T., Jiang, H., Kizer, L., Lieu, B., Melis, D., Moss, N., Regentin, R., Secrest, S., Tsuruta, H., Vazquez, R., Westblade, L.F., Xu, L., Yu, M., Zhang, Y., Zhao, L., Lievense, J., Covello, P.S., Keasling, J.D., Reiling, K.K., Renninger, N.S., Newman, J.D., 2013. High-level semi-synthetic production of the potent antimalarial artemisinin. Nature 496, 528–. doi:10.1038/nature12051

Pritchard, M.P., Ossetian, R., Li, D.N., Henderson, C.J., Burchell, B., Wolf, C.R., Friedberg, T., 1997. A general strategy for the expression of recombinant human cytochrome P450s in Escherichia coli using bacterial signal peptides: expression of CYP3A4, CYP2A6, and CYP2E1. Arch. Biochem. Biophys. 345, 342–354. doi:10.1006/abbi.1997.0265

Qi, X., Bakht, S., Qin, B., Leggett, M., Hemmings, A., Mellon, F., Eagles, J., Werck-Reichhart, D., Schaller, H., Lesot, A., Melton, R., Osbourn, A., 2006. A different function for a member of an ancient and highly conserved cytochrome P450 family: from essential sterols to plant defense. Proc. Natl. Acad. Sci. U.S.A. 103, 18848–18853. doi:10.1073/pnas.0607849103

Silva-Rocha, R., Martínez-García, E., Calles, B., Chavarría, M., Arce-Rodríguez, A., Las Heras, de, A., Páez-Espino, A.D., Durante-Rodríguez, G., Kim, J., Nikel, P.I., Platero, R., De Lorenzo, V., 2013. The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes. Nucleic Acids Res. 41, D666–75. doi:10.1093/nar/gks1119

Sonnhammer, E.L., Heijne, von, G., Krogh, A., 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6, 175–182.

Sonoda, Y., Newstead, S., Hu, N.-J., Alguel, Y., Nji, E., Beis, K., Yashiro, S., Lee, C., Leung, J., Cameron, A.D., Byrne, B., Iwata, S., Drew, D., 2011. Benchmarking Membrane Protein Detergent Stability for Improving Throughput of High-Resolution X-ray Structures. Structure 19, 17–25. doi:10.1016/j.str.2010.12.001

Studier, F.W., 1991. Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. J. Mol. Biol. 219, 37–44.

Studier, F.W., 2005. Protein production by auto-induction in high-density shaking cultures. Protein Expression and Purification 41, 207–234. doi:10.1016/j.pep.2005.01.016

Sudhamsu, J., Kabir, M., Airola, M.V., Patel, B.A., Yeh, S.-R., Rousseau, D.L., Crane, B.R., 2010. Co-expression of ferrochelatase allows for complete heme incorporation into recombinant proteins produced in E. coli. Protein Expression and Purification 73, 78–82. doi:10.1016/j.pep.2010.03.010

Wadsäter, M., Laursen, T., Singha, A., Hatzakis, N.S., Stamou, D., Barker, R., Mortensen, K., Feidenhans'l, R., Møller, B.L., Cárdenas, M., 2012. Monitoring shifts in the conformation equilibrium of the membrane protein cytochrome P450 reductase (POR) in nanodiscs. Journal of Biological Chemistry 287, 34596–34603. doi:10.1074/jbc.M112.400085

Wagner, S., Klepsch, M.M., Schlegel, S., Appel, A., Draheim, R., Tarry, M., Hogbom, M., Van Wijk, K.J., Slotboom, D.J., Persson, J.O., de Gier, J.W., 2008. Tuning Escherichia coli for membrane protein overexpression. Proc. Natl. Acad. Sci. U.S.A. 105, 14371–14376. doi:10.1073/pnas.0804090105

Waldo, G.S., Standish, B.M., Berendzen, J., Terwilliger, T.C., 1999. Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. 17, 691–695. doi:10.1038/10904

Williams, P.A., Cosme, J., Sridhar, V., Johnson, E.F., McRee, D.E., 2000. Mammalian microsomal

cytochrome P450 monooxygenase: Structural adaptations for membrane binding and functional diversity. Molecular Cell 5, 121–131.

Zelasko, S., Palaria, A., Das, A., 2013. Optimizations to achieve high-level expression of cytochrome P450 proteins using Escherichia coli expression systems. Protein Expression and Purification 92, 77–87. doi:10.1016/j.pep.2013.07.017

Zhang, H., Im, S.-C., Waskell, L., 2007. Cytochrome b(5) increases the rate of product formation by cytochrome p450 2B4 and competes with cytochrome p450 reductase for a binding site on cytochrome p450 2B4. J. Biol. Chem. 282, 29766–29776. doi:10.1074/jbc.M703845200

CHAPTER 2

An expression tag toolbox for microbial production of medicinal cytochromes P450

An expression tag toolbox for microbial production of medicinal cytochromes P450

Dario Vazquez-Albacete[1], Ana Mafalda Cavaleiro[1], Ulla Christensen[1], Susanna Seppälä[1], Birger Lindberg Møller[3,4] and Morten H. H. Nørholm[1,4]

[1] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle allé 6, Hørsholm, Denmark; [3] Plant Biochemistry Laboratory, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C, Copenhagen, Denmark; [4] Center for Synthetic Biology: bioSYNergy, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C, Copenhagen, Denmark.

Address correspondence to MHHN, morno@biosustain.dtu.dk. Phone: +45 217-99184 Fax: +45-353-33300

ABSTRACT

Background: Membrane-associated Cytochromes P450 (P450s) are one of the most important enzyme families for biosynthesis of plant-derived medicinal compounds. However, the hydrophobic nature of P450s makes their use in robust cell factories a challenge.

Results: We explore a small library of N-terminal expression tag chimeras of the model plant P450 CYP79A1 in different *Escherichia coli* strains. Using a high-throughput screening platform based on C-terminal GFP fusions, we identify several highly expressing and robustly performing chimeric designs. Analysis of long-term cultures by flow cytometry showed homogeneous populations for some of the conditions. Three chimeric designs were chosen for a more complex combinatorial assembly of a multigene pathway consisting of two P450s and a redox partner. Cells expressing these recombinant enzymes catalysed the conversion of the substrate to highly different ratios of the intermediate and the final product of the pathway. Finally, the effect of a robustly performing expression tag was explored with a library of 49 different P450s from medicinal plants and nearly half of these were improved in expression by more than 2-fold.

Conclusion: We have explored the effect of N-terminal expression tags in a metabolic engineering scenario and with a large number of medicinal P450s. The developed toolbox serves as platform to tune P450 performance in microbial cells, thereby facilitating recombinant production of complex plant P450-derived biochemicals.

Keywords: cytochromes P450, terpenoids, membrane protein, gene expression, N-terminal tag, medicinal plants

Background

Cytochromes P450 (P450s) are ubiquitous in plants, where they are involved in the biosynthesis of a wide range of bioactive secondary metabolites. These enzymes have attracted scientific and commercial interest because they modify complex carbon backbones in a highly stereo- and regiospecific manner. Plant P450s are typically membrane-bound by an N-terminal transmembrane helix and depend on a properly incorporated heme group in the catalytic domain. These features make plant P450s a particular challenge when it comes to recombinant production. Through a myriad of tailoring reactions P450s catalyze the last steps that lead to functionalization of many relevant medicinal compounds (Podust and Sherman 2012). Terpenoids are an example of such naturally occurring compounds, which have shown promising pharmacological activity (King, Brown et al. 2014). Unfortunately terpenoids and similar specialized metabolites are often produced in scarce amounts *in planta* making direct extraction from feedstocks economically unsustainable. Taxol and artemisinin are prominent examples of terpenoids used in treatment of cancer and malaria, respectively; both compounds are produced by medicinal plants through complex metabolic pathways involving P450s (Chang, Eachus et al. 2007, Ajikumar, Xiao et al. 2010).

Heterologous production of eukaryotic proteins in microbes such as *Escherichia coli* often requires a tedious optimization process. Hydrophobic membrane proteins, such as P450s, are particularly challenging as they tend to form insoluble inclusion bodies (Schlegel, Klepsch et al. 2010). High-throughput screening methods allow for fast design-build-test cycles through identification of highly expressing constructs, and reporters such as the green fluorescent protein from *Aequorea victoria* or bacteriorhodopsin from *Haloarcula marismortui* can enable quick assessment of properly targeted integral membrane proteins (Drew, Lerch et al. 2006, Hsu, Yu et al. 2013).

A common strategy to overcome challenges with membrane protein production consists of *E. coli* strains equipped with promoters suitable for subtle tuning of expression. For example, the KRX, the Lemo21(DE3) or the Walker strains C41 and C43 are known for higher tolerance for protein production (Miroux and Walker 1996, Giacalone, Gentile et al. 2006, Wagner, Klepsch et al. 2008). In the KRX strain, a rhamnose-tunable promoter drives expression of a genome-integrated T7 RNA polymerase whereas in the Lemo21(DE3) strain heterologous gene transcription is tuned by rhamnose-titrated expression of the T7 RNA polymerase inhibitor T7 lysozyme (Schlegel, Rujas et al. 2013). The Walker strains were evolved to tolerate high-level expression of a toxic membrane protein (Miroux and Walker 1996).

In addition to engineered strains, a handful of strategies have been attempted to specifically facilitate P450 expression in bacterial systems. These include truncation or replacement of the native membrane anchor to

solubilize the enzyme, and replacement of the plant membrane-targeting signal with a peptide better recognized by the bacterial membrane translocation machinery (Sibbesen, Koch et al. 1995, Pritchard, Ossetian et al. 1997, Mergulhao, Summers et al. 2005, Smith, Sanders et al. 2007). A popular strategy employs the engineered N-terminal sequence from the bovine P450 17α-hydroxylase, which has been used to produce mammalian and plant P450s in microbes (Barnes, Arlotto et al. 1991, Bak, Kahn et al. 1998). For example the engineering of several N-terminal P450 anchors resulted in higher artemisinin-precursor production in *E. coli*, showing the positive impact of N-terminal modifications for the complex biosynthesis of a valuable chemical (Chang, Eachus et al. 2007).

Although a number of N-terminal modifications are available for P450 tailoring, the toolbox for doing so is still limited, especially considering the enormous diversity of plant P450s and complex metabolic pathways. Thus, expanding the number of N-terminal tags, benchmarking the performance of such standard biological parts and testing their compatibility with a large number of e.g. plant P450s sequences could be of great value.

Here, we explore a number of different N-terminal tags for microbial P450 production using the well-characterized *Sorghum bicolor* CYP79A1 as a model plant P450 (Nielsen and Moller 2000). Further, we explore the effect of these peptides in different microbial strains and on a single cell level by flow cytometry. Finally, we demonstrate the use of the N-terminal tag toolbox for subtle expression of a more complex multigene P450 pathway and with a larger library of 49 different medicinal P450s.

Results

Expression analysis of a library of N-terminal tags of *Sb*CYP79A1

We constructed a synthetic N-terminal tag library for the plant cytochrome P450 CYP79A1. The N-terminal tags were cloned as CYP79A1 fusions in a pET28a(+) expression vector. The vector contains a T7 promoter upstream of the gene, a TEV protease cleavage site at the 3´end of the gene followed by a GFP folding reporter and an octahistidine tag (Norholm, Toddo et al. 2013). The N-terminal tags were chosen based on their topologies and expression profiles determined in other studies (Daley, Rapp et al. 2005, Sletta, Tondervik et al. 2007, Lee, Velmurugan et al. 2013), membrane localization in *E. coli* (Gotzke, Muheim et al. 2015) or relaxation of mRNA folding energy (Kudla, Murray et al. 2009). The N-terminal tag-types fall in five categories (Fig. 1a): 1) bacterial membrane anchors with the C-terminal facing the cytoplasmic side (C-in; SohB, LepA, ZipA and YcjF), 2) proteins with the C-terminal facing the periplasmic side (C-out; YfdG,

YhaH, YmcD, PpiD and YfgM), 3) signal peptides (MalE, OmpA, DsbA and NlpA_as carrying a lipid anchoring sequence (as) and NlpA_ns without lipid anchoring sequence), 4) transporters (GltS and TyrP) and 5) a previously described expression-enhancing peptide (28-tag) (Norholm, Toddo et al. 2013). The C-in tags and transporters replaced the transmembrane segment of the P450; whereas in the other constructs the P450 transmembrane segment was left intact.

The N-terminal tag chimeric constructs of CYP79A1 were transformed into the KRX *E. coli* strain that harbor a T7 RNA polymerase under the control of the tunable rhamnose promoter. Clones were grown on TB medium and induced with both rhamnose and IPTG for 24 hours. Single cell and whole cell fluorescence were subsequently measured as indication of the presence of CYP79A1 chimeras. The data shows no fluorescence of the wild type *cyp79a1* gene fusion in *E. coli*, whereas all N-terminal tagged chimeras were produced to a detectable level (Fig. 1b). Most of the chimeras under the C-in category, which required removal of the CYP79A1 transmembrane segment, exhibited the strongest fluorescent signal together with the 28-tag chimera. Yields were estimated based on levels of GFP and were up to 40 mg/L for the best performing constructs. Chimeras based on the C-out category and signal peptides were less fluorescent than the C-in constructs, with the YfgM chimera performing best at 15 mg GFP/L. The presence of full length chimeras in membrane fractions was confirmed for highest expressed constructs by cell fractionation and SDS-PAGE (Fig S1).

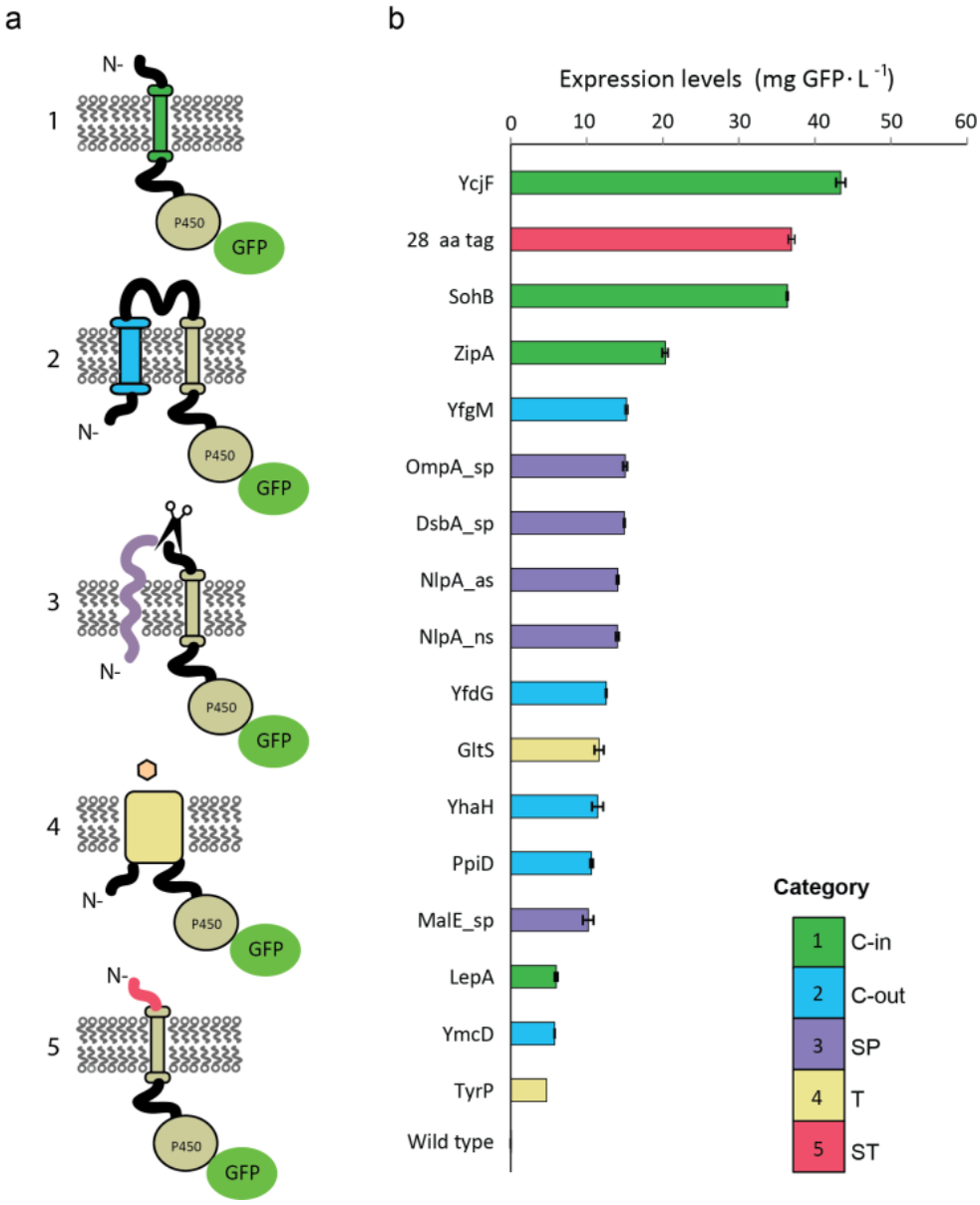


Fig. 1 Expression screening of N-terminally tagged CYP79A1 chimeras in the *Escherichia coli* KRX strain.. **a** Illustration of the protein architectures included in the N-terminal tag library with colour coding. Five categories of N-terminal tags are shown in different colours; Green for C-in architectures, blue for C-out architectures, purple for signal peptides, beige for transporters and red for the short 28-tag sequence. **b** Expression levels of N-terminal-tagged chimeras and wild type CYP79A1 (no tag). Expression was induced with IPTG and rhamnose in exponential phase for 24 h and expression levels estimated by whole cell fluorescence in a plate reader. Data represents the average of three biological replicates with standard deviations.

Next, the CYP79A1 chimeras were tested in the *E. coli* B strain Lemo21(DE3) (Wagner, Klepsch et al. 2008). Furthermore, to find the best expression conditions for each chimera considering maximum GFP

signal and homogeneity of cell populations, the cells were induced with three different rhamnose concentrations. As in KRX, the wild type *cyp79a1* gene was poorly expressed and expression was significantly improved by nearly all N-terminal tags (Fig 2a). While in the KRX strain the highest fluorescence was observed for C-in chimeras, in the Lemo21(DE3) strain, the highest signal was obtained for the PpiD tag (C-out) in the absence of rhamnose, followed by the signal peptide chimeras and the 28-tag at 0.5 mM rhamnose, with values ranging from 80 mg GFP/L to 100 mg GFP/L (Fig 2a). Different constructs responded differently to rhamnose induction. As judged from GFP-levels, cells expressing signal peptide- and 28-tag chimeras performed best with the lowest concentration inducer (0.5 mM rhamnose), whereas cells expressing most of the other chimeras responded negatively to any presence or increase in rhamnose concentration, corroborating the notion that inducer levels should be titrated for each construct (Fig 2a). For the signal peptide- and 28-tag-chimeras, in the absence of rhamnose, GFP fluorescence was lost after 24 hours, whereas the addition of 0.5 mM rhamnose appears not only to give higher expression, but, importantly, fluorescence is maintained after 24 hours in homogenous populations. This demonstrates the value of subtle expression titration for increasing phenotype stability.

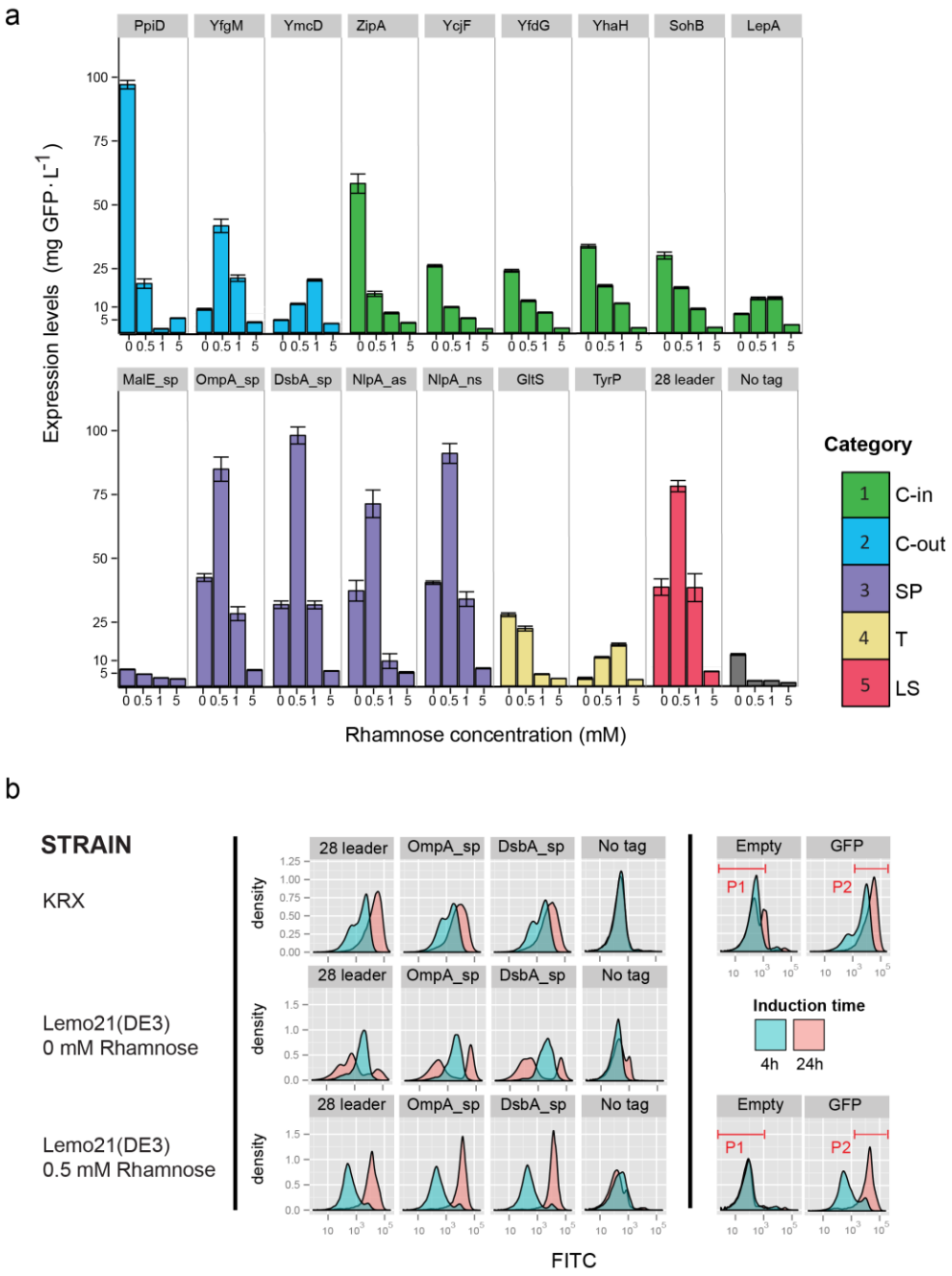


Fig. 2 Expression screening of N-terminally tagged CYP79A1 chimeras in the *E. coli* strains Lemo21(DE3) and KRX analysed by whole-cell fluorescence and flow cytometry. **a** Expression levels of N-terminal-tagged chimeras and wild type CYP79A1 (non tagged) in the Lemo21(DE3) strain. Gene expression was induced with IPTG and rhamnose (0, 0.5, 1 or 5 mM) in exponential phase for 24 h and whole cell fluorescence was measured in a plate reader. Data represents the average of three biological replicates with standard deviations. Five categories of N-terminal tags are shown in different colours; Green for C-in architectures, blue for C-out architectures, purple for signal peptides, beige for transporters and red for the short 28-tag sequence. **b** Single cell fluorescence measurements by flow cytometry of selected CYP79A1 chimeras

expressed in KRX and Lemo21(DE3).. Expression was induced for four or 24 h as described above and different constructs were transferred to PBS and fluorescence analysed in the FITC channel. Negative and positive expression thresholds were determined with control cells containing the empty vector (P1) and cells expressing soluble GFP (P2), respectively.

Taken together, these results showcase the highly different performance of different N-terminal tags on balancing expression of a model P450. Signal peptides and the 28-tag showed high fluorescence of P450 chimeras in both strains. Flow cytometry analysis showed loss of long term expression with some chimeric constructs and subtle expression tuning helped maximizing and stabilizing the phenotype.

Multigene pathway activity

Dhurrin is a natural compound synthesized from the amino acid tyrosine and released by plants to deter herbivores from feeding on leaves. The first two steps in the dhurrin pathway of *Sorghum bicolor* are catalyzed by the cytochromes P450 CYP79A1 and CYP71E1, respectively. The two intermediates generated by each enzyme, (*E*)-*p*-hydroxyphenylacetaldoxime and (*E*)-p-hydroxymandelonitrile, can be easily detected and separated from the substrate tyrosine by thin layer chromatography (TLC) (Kahn, Fahrendorf et al. 1999). The catalytic reaction is assisted by the electron donor cytochrome P450 reductase (*Sb*CPR2b) (Laursen, Jensen et al. 2011). Seeing the robust performance of the OmpA, DsbA and 28-tagged CYP79A1 chimeras, we expanded our studies by systematically combining chimeric versions of the dhurrin pathway P450s in the same expression vector as two separate transcriptional units consisting of a T7P promoter combined with a T1 terminator for the CYP79A1 chimeras or a T7 terminator for CYP71E1 chimeras (Fig 3a and S2a). The vectors harboring the P450s and the partnering reductase (CPR) ORFs were co-transformed and expressed for 24 h in the KRX strain. The activity of the pathway was subsequently assayed by addition of radioactively labelled tyrosine to cultures normalized to the same optical density units (ODU). The two oximes produced by CYP79A1 and CYP71E1 were extracted from the reaction with ethyl acetate and analyzed by TLC (Fig 3b, Fig. S2b). Cells producing a combination of 28-tagged *Sb*CYP79A1 and wild type *Sb*CYP71E1 produced only the first intermediate, (*E*)-*p*-hydroxyphenylacetaldoxime at a high level (Fig. 3b). When both P450s are N-terminally tagged, the (E)-p-hydroxymandelonitrile is produced and the combination of a 28-tagged *Sb*CYP79A1 and an OmpA-tagged *Sb*CYP71E1 yielded the highest level of the final product of the pathway. As expected from the low expression level, no combinations with the wild type cytochrome *cyp79a1* generated any product. These results demonstrate the usefulness of an extended N-terminal tag toolbox for cell factory design.

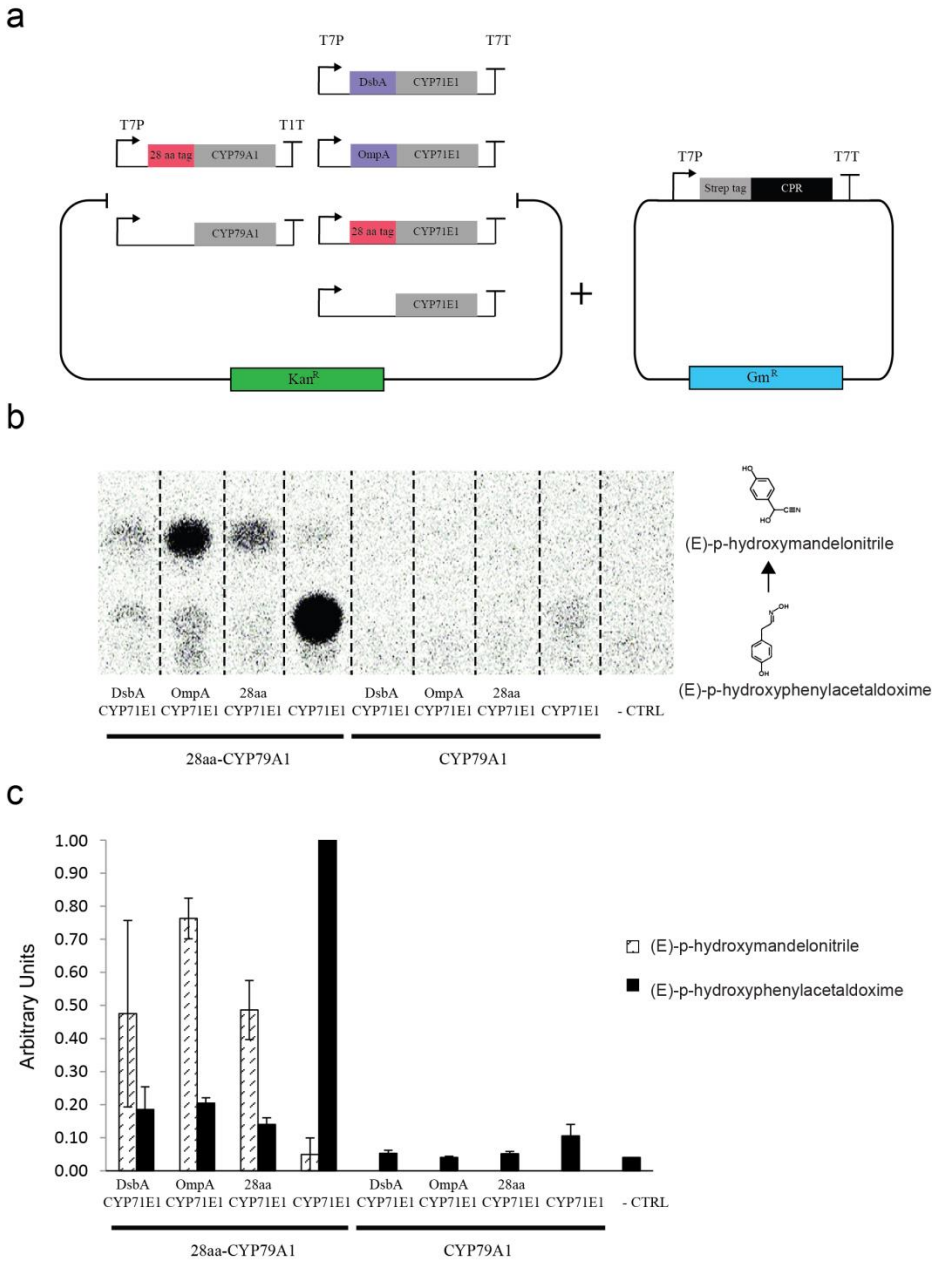


Fig. 3 Activity of the dhurrin pathway assembled and assayed in the *E. coli* KRX strain. **a** Left side: Schematic representation of the dhurrin pathway assembled as two separate transcriptional units in the pET28a(+) vector. Combinations of the wild type and N-terminal-tagged cytochromes were assembled into the pET28 vector containing two T7 promoters (T7P) and two different terminators following each ORF. The constructs were co-transformed in *E. coli* with a cytochrome P450 reductase (CPR) construct illustrated on the right side. **b** Analysis of activity of the dhurrin pathway P450s by Thin-Layer-Chromatography (TLC). Radioactive spots observed in the lower part of the TLC correspond to (*E*)-*p*-hydroxyphenylacetaldoxime, the enzymatic product of CYP79A1, whereas the spots observed in the upper part correspond to (*E*)-p-hydroxymandelonitrile, produced by CYP71E1. Chemical strcutures are illustrated

on the right side of the TLC. Cells were grown in TB media and expression of the multigene pathway was induced for 24 h. Cells were adjusted to the same ODU, fed with [U-^{14}C]tyrosine and incubated for 1 h. Enzymatic products were extracted from the reaction with ethyl acetate and loaded onto a TLC silica plate. Radioactive products were visualized using phosphor imaging. **c** Relative levels of oximes from two biological replicates were quantified in each TLC plate using ImageJ and normalized to the most intense spot.

Expression of a library of tagged medicinal P450s

To test the broader applicability of the N-terminal tagging strategy, we next synthesized a library of 49 codon harmonized P450 genes derived from the medicinal plants *Euphorbia peplus* (*Ep*), *Coleus forskohlii* (*Cf*), *Thapsia garganica* (*Tg*), *Thapisa villosa* (*Tv*), *Tripterygium willfordii* (*Tw*) and the insect *Heliconius Melpomene* (*Hm*) that has metabolic pathways and P450s that are related to plant enzymes (Chauhan, Jones et al. 2013). The genes were produced with and without an N-terminal 28-tag, all fused to a C-terminal GFP as described above. The fold change in fluorescence between the wild type and chimeric P450s was used to estimate expression level differences. The 28-tag improved expression of 68% of the P450 genes and expression of 44% of the library was increased two-fold or more with the 28-tag.

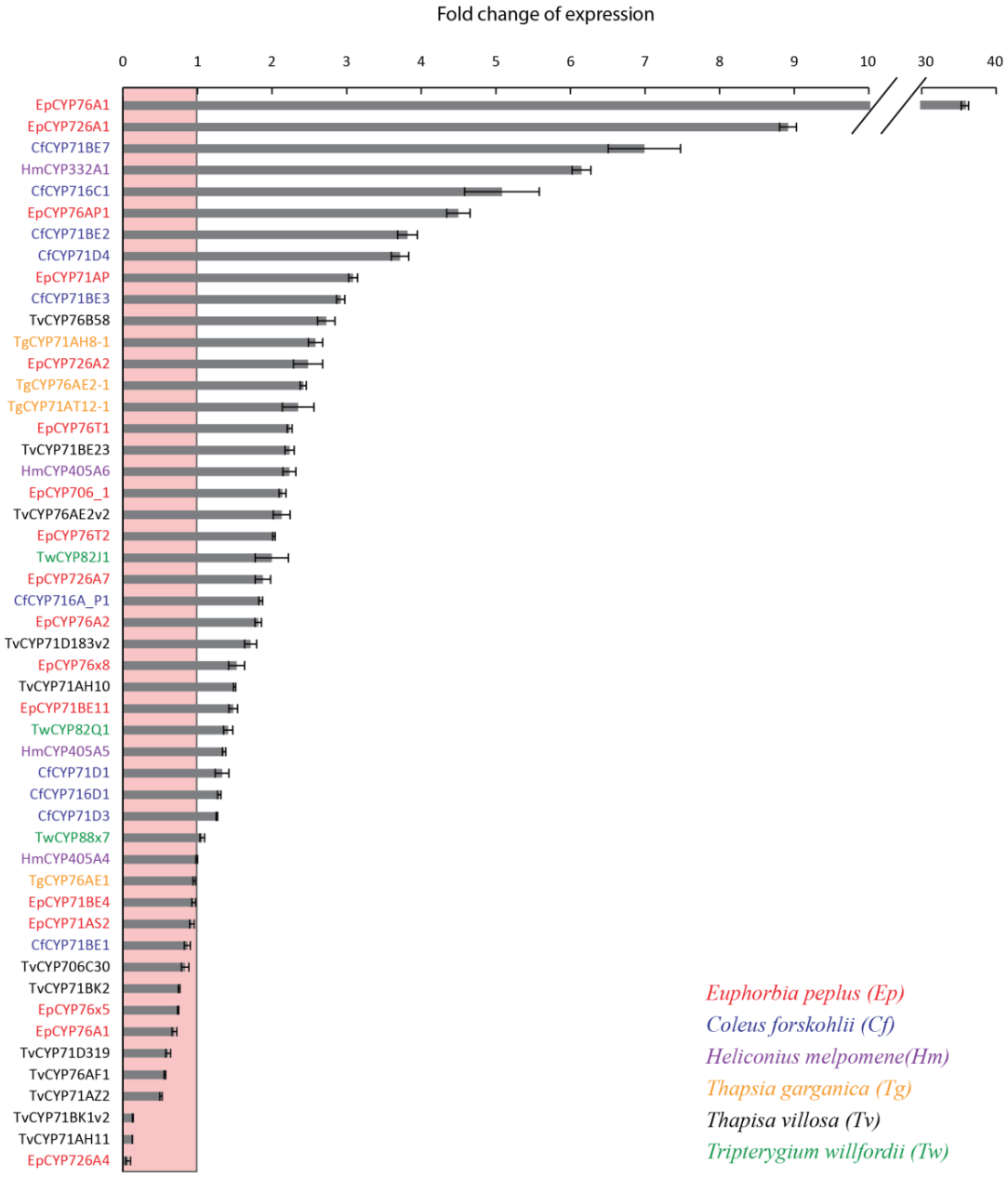


Fig. 4 Average fold change in expression levels between wild type and 28-tagged P450-encoding genes estimated by GFP fluorescence in *E. coli* KRX. Expression was induced with IPTG and rhamnose in exponential phase for 24 h and expression levels estimated by whole cell fluorescence in a plate reader. Data represents the average of three biological replicates with standard deviations.

Discussion

Membrane-associated cytochromes P450 are important biocatalysts in the biosynthesis of natural compounds such as terpenoids (Podust and Sherman 2012). A major bottleneck for the production of functionalized terpenoids in microbial hosts is robust expression (Chang and Keasling 2006). Here, we assayed the performance and stability of *E. coli* cultures expressing N-terminally modified plant P450 enzymes. Using a GFP-based platform that allows simple assessment of expression levels, we show that modifying the plant enzymes with signal peptides or protein tags derived from the microbial host significantly increase levels of the chimeric proteins, suggesting that host-tailored modifications of the target protein may be key to heterologous protein production. We also specifically verified the beneficial effect on gene expression of a 28 amino acid tag that has been shown to boost expression, possibly by relaxing mRNA structure (Kudla, Murray et al. 2009). Further, we compared the performance of the two different *E. coli strains* KRX and Lemo21(DE3) (Miroux and Walker 1996, Giacalone, Gentile et al. 2006, Wagner, Klepsch et al. 2008) - both strains allow for tunable gene expression and have previously shown promise for the production of challenging proteins. Whole cell fluorescence varied between the strains, with Lemo21(DE3) strain outperforming KRX in terms of total estimated protein yield. However, this pattern could change upon subtle tuning of T7 polymerase expression with rhamnose in KRX.

Different N-terminal tag chimerase showed subtle expression differences in the two strains and with different rhamnose levels in Lemo21(DE3). For example, using flow cytometry, with the signal peptide base constructs in the Lemo21(DE3) strain, 0.5 mM rhamnose stabilized the homogeneity and maintained expression at high levels. In contrast, absence of rhamnose led to loss of fluorescence in several cultures over time. This is likely due to over-expression of the P450 saturating the Sec translocon machinery, which may result in protein aggregation (Klepsch, Persson et al. 2011). The two rhamnose-based expression strains had no or minimal effects on increasing fluorescence of the wild type gene *cyp79a1* fusion. Therefore it is clear that an N-terminal tag library provides important tools to unblock expression bottlenecks in *E. coli*.

We further demonstrated the impact of combinatorial chimeric constructs on pathway optimization and microbial production of the plant-derived chemical dhurrin. Two P450s participate in the formation of dhurrin from tyrosine and the P450s receive electrons from an associated oxidoreductase. Using the best overall performing expression tags we assembled different combinations of the dhurrin pathway as separate transcriptional units in *E. coli*, and assayed for the formation of the intermediate (*E*)-*p*-hydroxyphenylacetaldoxime and product (*E*)-p-hydroxymandelonitrile. Only when the P450s were tagged, product formation was detected, but the different combinations produced highly different amounts of the final product of the pathway. This emphasizes the usefulness of an extended toolbox of different N-terminal tags, to serve as platform for biosynthetic pathway engineering.

We further analyzed the effect of the 28-tag on a library of 49 P450 genes from five medicinal plant species and one insect. These coding sequences were harmonized for codon usage in *E. coli*, but many factors determine the success of codon optimization strategies (Menzella 2011). Our results show that a large number of P450s from different species are positively affected by N-terminal tags even when they are all codon optimized for the same organism. The most dramatic improvement was observed for *Ep*CYP76A1, which has not previously been expressed in a microbial host and other members of this family were also well-expressed. This is exciting, because *Arabidopsis thaliana* and *Catharantus roseus* homologs of the CYP76 family have been associated with geraniol and hydroxygeraniol hydroxylase activities - key enzymes in the early steps in the biosynthesis of valuable chemicals such as the anticancer molecule vinblastine (Hofer, Dong et al. 2013). Further, we observed a 9-fold expression increase for *Ep*CYP726A1, which is involved in the biosynthesis of important diterpenoids such as ingenol, a commercially available medicinal compound currently produced by a complex 9-step chemical reaction (Jorgensen, McKerrall et al. 2013, King, Brown et al. 2014). Future production of these medicinal compounds in bacteria may be assisted by the aid of the N-terminal tag toolbox described here.

Conclusion

In this study we have found and characterized N-terminal tags that allow or fine-tune expression of functionally active plant P450s in multigene biosynthetic pathways. We further demonstrate the use of these tags in two commonly used *E. coli* protein production strains, test their performance on the single-cell level and a with large medicinal P450 library. These results provide a framework for the broad application of N-terminal tags in the design of microbial cell factories.

Methods

Strains and growth conditions

For cloning NEB® 5-alpha Competent *E. coli* cells (New England Biolabs, MA, USA) were transformed and propagated in standard LB media as described by the manufacturer. Chemically competent *E. coli* strains KRX (Promega, Madison, USA) and One Shot® BL21(DE3) (Invitrogen, CA, USA) bearing the pLemo plasmid (referred as strain Lemo21(DE3)), were transformed with the corresponding pET28a(+)-derived constructs (Wagner, Klepsch et al. 2008, Mirzadeh, Martinez et al. 2015). Cells were grown in Terrific Broth (TB) and, for Lemo21(DE3), 0.5, 1 and 5 mM of rhamnose (Sigma-Aldrich, St. Louis, USA) was supplemented in the medium. Three biological replicates of each construct were inoculated in TB media

supplemented with 50 µg/mL Kanamycin and 25 µg/mL Chloramphenicol and grown over-night. Expression cultures were initiated at a final OD of 0.05 in 200 µl of fresh TB medium in 96-well cell culture microtiter plates. Absorbance measurements were carried out at 600 nm in a SynergyMx, SMATLD plate reader (BioTek, Winooski, USA). Growth was performed for approximately 2 h at 30°C, 250 rpm in an Innova®44R incubator shaker system (5 cm orbital shaking) (New Brunswick Scientific, Eppendorf, USA) upon reaching 0.5 optical density. Protein expression was induced with IPTG (isopropyl ß-D-1-thiogalactopyranoside, dioxane free, Thermo Scientific, Waltham, USA) at a final concentration of 0.4 mM and rhamnose 5 mM for KRX strain (Sigma-Aldrich, MO, USA). Expression cultures were grown at 25°C for 24 h. shaking.

PCR and uracil excision cloning

Uracil excision cloning with PCR-amplified fragments was used to assemble all DNA constructs as described previously (Cavaleiro, Kim et al. 2015). All oligonucleotides used are listed in Table S1. PCR products were purified from 1% agarose gels using NucleoSpin Gel and PCR Clean-up (Macherey-Nagel, Düren, Germany). Purified PCR products were mixed with 1 µL USER^{TM} enzyme (New England BioLabs, Ipswich, USA) for 30 min at 37°C and at the *Tm* of the respective primer overhangs for 15 min. DNA was quantified in Nanodrop spectrophotometer 2000 instrument (Thermo Scientific, Waltham, USA). Transformations were carried out in chemically competent NEB5α cells with 5 µL of the uracil excision reaction. Transformants were screened by colony PCR with OneTaq 2X Master mix (New England BioLabs, Ipswich, USA). for plasmid isolation with a QIAprep Spin Miniprep Kit (Qiagen) and sequenced according to the service supplier instructions (Eurofins Genomics, Ebersberg, Germany). All resulting plasmid constructs are listed in Table S2.

Generation of N-terminally tagged P450 libraries

A library of different synthetic N-terminal tags for the cytochrome *Sb*CYP79A1 was constructed by uracil excision cloning as described above. A pET28a(+) expression vector bearing the native sequence of the *Sorghum bicolor* cytochrome *Sb*CYP79A1, a TEV protease cleavage site, GFP folding reporter and His-tag tail (*pET-CYP79A1*) was used as template to insert the different N-terminal tags (Drew, von Heijne et al. 2001). A version of the same plasmid with the *Sb*CYP79A1 lacking the amino acids one to 35 (t*Sb*CYP79A1) was also prepared. The genes *ppiD, yfgM, yfdG, yhaH, ymcD, gltS, tyrP* and the 28-tag sequence were amplified by PCR with the oligonucleotides (L4-L17) from *E. coli* MG1665 genome and cloned by two-fragment uracil excision cloning into the *pET-CYP79A1* backbone linearized with

oligonucleotides L2 and L3. The predicted transmembrane region (residues 1-39) of *zipA* and the two predicted transmembrane helices of the *ycjF* (residues 1-147) were cloned the same way into the *pET-tCYP79A1* plasmid linearized with oligonucleotides L1 and L3. Finally, signal peptides from *dsbA*, *ompA*, *malE* and *nlpA* with the lipid anchoring sequence CDQSSS (*nlpA*_as) and without this sequence (*nlpA*_ns) were cloned into the *pET-CYP79A1* plasmid by one-fragment uracil excision cloning using the oligonucleotides L22 to L34. Predictions of transmembrane regions from endogenous *E. coli* membrane proteins were performed with TMHMM v. 2.0 and TOPCONS (Hessa, Meindl-Beinker et al. 2007, Tsirigos, Peters et al. 2015). Signal peptides were predicted using SignalP 4.0 server (Petersen, Brunak et al. 2011).

Additionally, a library of 49 codon optimized P450-coding genes from *Euphorbia peplus (Ep)*, *Coleus forskohlii (Cf)*, *Heliconius Melpomene(Hm)*, *Thapsia garganica (Tg)*, *Thapisa villosa (Tv)* and *Tripterygium willfordii (Tw)* was purchased from GeneArt™ (Thermo Fisher Scientific, Wilmington, USA). ORFs were inserted in the pET28a(+) vector flanked by restriction sites *Xho*I and *Bam*HI (Table S3). Oligonucleotides L49 and L50 used to PCR amplify the pET28a(+) backbone had 5´-end *Xho*I and *Bam*HI restriction sites. Plasmids were digested with Fast Digest™ *Xho*I and *Bam*HI (Thermo Fisher Scientific, Wilmington, USA) for 20 min at 37°C, 5 min at 80°C and 10 min at 4°C. Ligation was performed with T4 DNA ligase (Thermo Fisher Scientific, Wilmington, USA) following specifications of the manufacturer. Mixtures were kept on ice prior to transformation.

Dhurrin pathway assembly

The N-terminal tags *dsbA*, *ompA* and the 28-tag sequence versions of the cytochrome *Sb*CYP71E1 were cloned as described for the *Sb*CYP79A1 in section 2.2 with oligonucleotides L35 to L40. For assembly of the multigene pathway, two-fragment uracil excision was performed; tagged and non-tagged versions of the two cytochromes without the GFP moiety were cloned in all possible combinations as two separate transcriptional units with the oligonucleotides L41 to L44. See list of resulting constructs (Table S2).

A cytochrome reductase from *Sorghum bicolor* (*CPR*) containing a Strep-HRV3C tag in the 5´end was also cloned by two-fragment uracil excision into the pET28a(+) vector from which the GFP folding reporter was removed. The origin of replication and the antibiotic resistance marker were swapped by the corresponding pSEVA63 parts to avoid plasmid incompatibility (Durante-Rodriguez, de Lorenzo et al. 2014).

GFP measurements and expression levels

Whole cell fluorescence was measured in a SynergyMx SMATLD plate reader (BioTek, Winooski, USA) from 96-well cell culture plates within the 485-512 nm range. To estimate expression levels, purified GFP was used as standard, diluted in cultured cells transformed with the empty vector (negative control) in order to simulate the quenching effect as described previously (Mirzadeh, Martinez et al. 2015). Single cell measurements were also performed in a BD LSRFortessa™ flow cytometer equipped with HTS module (Becton, Dickinson and Company, USA). Cells were diluted 1/200 in PBS, incubated for 15 min at room temperature and green fluorescence was collected in the FITC channel (488 nm emission wavelength). Expression thresholds were determined with controls consisting of cells containing the empty vector and soluble GFP expressing cells for both tested *E. coli* strains. Flow cytometry data was exported in FCS files and analysed using the *flowViz* and *flowcore* packages with *R* software (Sarkar, Le Meur et al. 2008, Hahne, LeMeur et al. 2009) (http://www.R-project.org/)

Preparation of *E. coli* membrane fractions

In order to recover the overexpressed membrane-associated cytochromes, *E. coli* membranes were isolated as previously described (Drew, Lerch et al. 2006). Briefly, expression was carried out using 50 ml in flasks and harvested after 24 h as for the previous experiments. Cells were lysed by mechanical disruption using an Emulsiflex C-50 (Avestin Europe GmbH) instrument at 15,000-20,000 p.s.i. at 4°C. Unbroken cells were removed by centrifugation at 8,000 x*g*, 10 min at 4˚C and the supernatant collected for further fractionation. Membranes were separated by ultracentrifugation at 150.000 x*g*, 45 min, 4°C and recovered from the pellet in 20 mM Tris buffer pH=7.8, 20% glycerol. Protein concentration was determined using BCA protein assay (Thermo Scientific, USA) prior to storage at -80˚C.

SDS-PAGE analysis

Membrane proteins were analyzed by standard SDS-PAGE using 10% acrylamide gels and loading 100 µg of each sample. Fluorescence proteins were detected in a G:BOX Bioimager UV-table (Syngene, Cambridge, UK).

Activity assay of dhurrin pathway cytochrome P450 enzymes and detection by Thin Layer Chromatography (TLC)

The activity of the P450s in the dhurrin pathway was analyzed qualitatively as previously reported with minor modifications (Kahn, Fahrendorf et al. 1999). *E. coli* cells transformed with the two-cytochrome pathway and the CPR were cultured accordingly in 1 ml volume. After 24 h expression, absorbance was measured; cells were washed twice with 50 mM KPi buffer (Potassium phosphate) and adjusted to the same optical density units (ODU). The pathway activity was assayed *in vivo* in a reaction volume of 15 µL consisting of 10 µL of cell suspension (ODU=1), 5 mM NADPH, 0.1 mM L-tyrosine (Sigma-Aldrich, St.

Louis, USA), 1.5 µL of [U-^{14}C]tyrosine (0.05 mCi, 482 mCi/mmol) and adjusted with 50 mM KPi buffer. The reaction was performed for 60 min at 30°C and 700 rpm in a benchtop shaker (Grant Instruments, Cambridge, UK). The products of the pathway, (*E*)-*p*-hydroxyphenylacetaldoxime and p-hydroxymandelonitrile (oximes), were extracted from the reaction by addition of one volume of ethyl acetate (Sigma-aldrich, MO, USA). The mixture was centrifuged to remove cell debries at 16,000 xg for 10 min and the organic phase transferred to a silica gel plate (Sigma-Aldrich, St. Louis, USA). The extract was eluted in a glass chamber with a mobile phase composed by toluene, ethyl acetate and methanol (30:8:1). Radioactive products were visualized in a Cyclone Plus Phosphor Imager instrument (Perkin Elmer, Massachusetts, USA). Relative intensity of radioactive oximes was determined by *ImageJ* (U. S. National Institutes of Health, Bethesda, Maryland, USA, http://imagej.nih.gov/ij/).

List of abbreviations

P450: Cytochrome P450s

GFP: green fluorescent protein

CYP79A1: Cytochrome P450 CYP79A1

CYP79E1: Cytochrome P450 CYP79E1

CPR: Cytochrome P450 Reductase

28-tag: short 28 amino acid tag

TLC: Thin Layer Chromatography

Funding

This work was supported by The Novo Nordisk Foundation and a PhD grant from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme [FP7-People-2012-ITN], under grant agreement No. 317058, "BACTORY". SS is the recipient of VILLUM Foundation's Young Investigator Programme grant VKR023128.

Authors' contributions

Dario Vazquez designed the experiments, carried out cloning of N-terminal tag library, optimized the expression platform and analyzed membrane proteins. Mafalda Cavaleiro designed cloning strategy of P450 libraries, cloned a significant number of constructs and participated in the expression experiments. Ulla Chirstensen designed the initial expression platform and provided guidance during the study. Susanna Seppälä designed N-terminal tag chimeras and provided guidance during the study. Birger Lindberg Møller provided insights into biochemistry and diversity of P450 enzymes. Morten H.H. Nørholm proposed, supervised and provided unique guidance of the study.

Acknowledgements

Björn Hamberger, Søren Bak and Henrik Toft Simonsen are thanked for assistance with selecting P450s for the expression library. Acknowledgement is given to João Cardoso for helping with the R code.

Supplementary figures

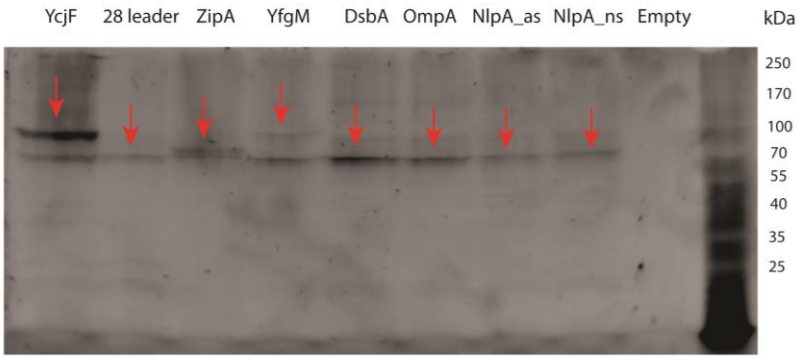


Fig. S1 SDS-PAGE nn-gel fluorescence analysis of membrane fractions from N-terminally tagged CYP79A1 chimeras and empty vector-containing. Red arrows indicate the likely full-length CYP79A1-GFP chimeric products. Expected sizes for CYP79A1 chimeras; YcjF (104 kDa), 28-tag (90 kDa), ZipA (93.4 kDa), YfgM (114 kDa), DsbA (93.9 kDa), OmpA (93.8 kDa), NlpA_as (94.7 kDa), NlpA_ns (93.1 kDa). Membrane fractions were prepared in the KRX strain after 24 h incubation. Cells were harvested, washed and mechanically disrupted. Membranes were recovered after ultracentrifuge fractionation. 100 µg of membrane protein were loaded on a 10% polyacrylamide gel for analysis.

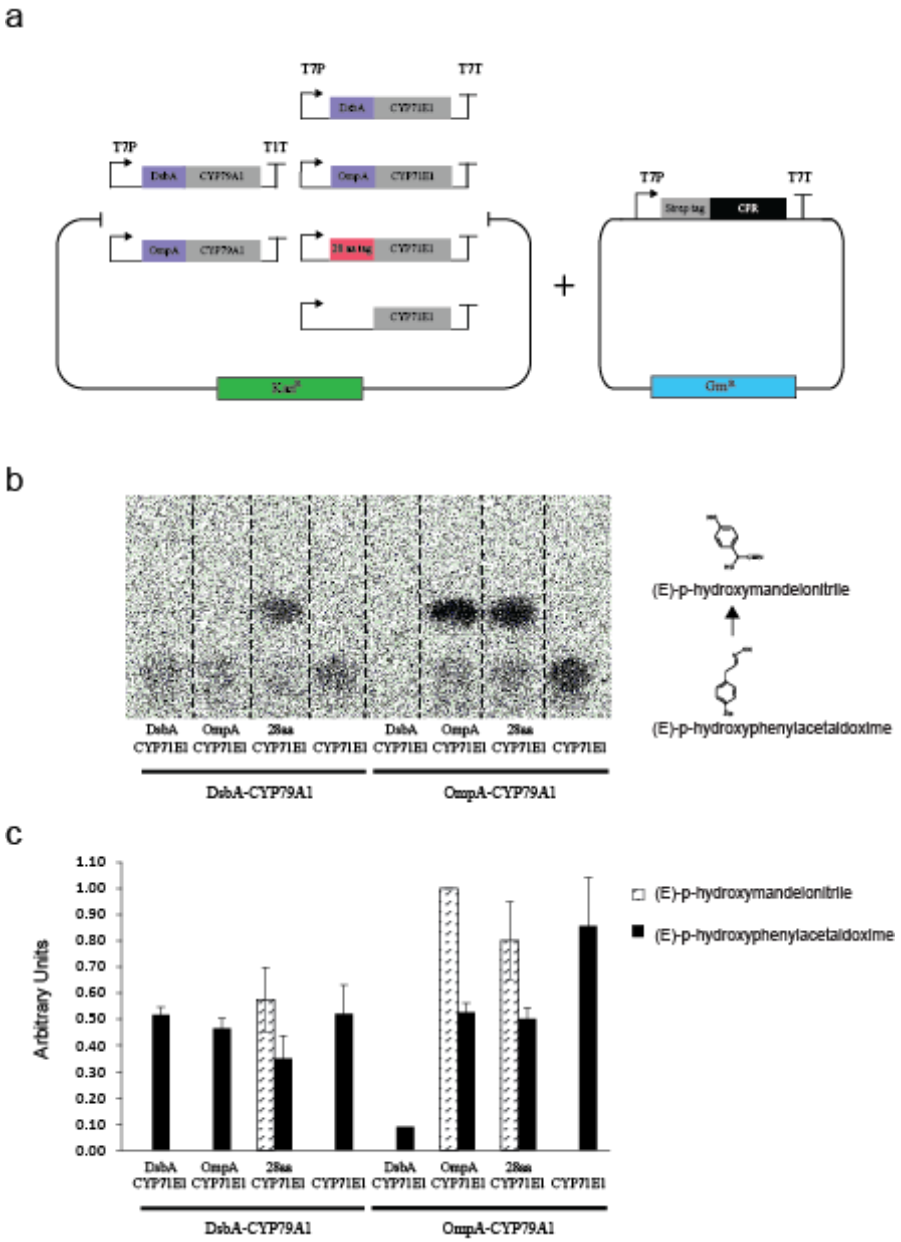


Fig. S2 Activity of the dhurrin pathway assembled and assayed in the *E. coli* KRX strain. **a** Left side: Schematic representation of the dhurrin pathway assembled as two separate transcriptional units in the pET28a(+) vector. Combinations of the wild type and N-terminal-tagged cytochromes were assembled into the pET28 vector containing two T7 promoters (T7P) and two different terminators following each ORF. The constructs were co-transformed in *E. coli* with a cytochrome P450 reductase (CPR) construct illustrated on the right side. **b** Analysis of activity of the dhurrin pathway P450s by Thin-Layer-Chromatography (TLC). Radioactive spots observed in the lower part of the TLC correspond to (*E*)-*p*-hydroxyphenylacetaldoxime, the enzymatic product of CYP79A1, whereas the spots observed in the upper part correspond to (*E*)-p-hydroxymandelonitrile, produced by CYP71E1. Chemical strcutures are illustrated on the right side of the TLC. Cells were grown in TB media and expression of the multigene pathway was

induced for 24 h. Cells were adjusted to the same ODU, fed with [U-^{14}C]tyrosine and incubated for 1 h. Enzymatic products were extracted from the reaction with ethyl acetate and loaded onto a TLC silica plate. Radioactive products were visualized using phosphor imaging. **c** Relative levels of oximes from two biological replicates were quantified in each TLC plate using ImageJ and normalized to the most intense spot.

References

1. Podust L, Sherman D: **Diversity of P450 enzymes in the biosynthesis of natural products.** *Natural product reports* 2012, **29:**1251-1266.
2. King AJ, Brown GD, Gilday AD, Larson TR, Graham IA: **Production of bioactive diterpenoids in the euphorbiaceae depends on evolutionarily conserved gene clusters.** *Plant Cell* 2014, **26:**3286-3298.
3. Chang MC, Eachus RA, Trieu W, Ro DK, Keasling JD: **Engineering Escherichia coli for production of functionalized terpenoids using plant P450s.** *Nat Chem Biol* 2007, **3:**274-277.
4. Ajikumar PK, Xiao WH, Tyo KE, Wang Y, Simeon F, Leonard E, Mucha O, Phon TH, Pfeifer B, Stephanopoulos G: **Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli.** *Science* 2010, **330:**70-74.
5. Schlegel S, Klepsch M, Gialama D, Wickstrom D, Slotboom DJ, de Gier JW: **Revolutionizing membrane protein overexpression in bacteria.** *Microb Biotechnol* 2010, **3:**403-411.
6. Hsu MF, Yu TF, Chou CC, Fu HY, Yang CS, Wang AH: **Using Haloarcula marismortui bacteriorhodopsin as a fusion tag for enhancing and visible expression of integral membrane proteins in Escherichia coli.** *PLoS One* 2013, **8:**e56363.
7. Drew D, Lerch M, Kunji E, Slotboom DJ, de Gier JW: **Optimization of membrane protein overexpression and purification using GFP fusions.** *Nat Methods* 2006, **3:**303-313.
8. Wagner S, Klepsch MM, Schlegel S, Appel A, Draheim R, Tarry M, Hogbom M, van Wijk KJ, Slotboom DJ, Persson JO, de Gier JW: **Tuning Escherichia coli for membrane protein overexpression.** *Proc Natl Acad Sci U S A* 2008, **105:**14371-14376.
9. Giacalone M, Gentile A, Lovitt B, Berkley N, Gunderson C, Surber M: **Toxic protein expression in Escherichia coli using a rhamnose-based tightly regulated and tunable promoter system.** *BioTechniques* 2006, **40:**355-364.

10. Miroux B, Walker JE: **Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels.** *J Mol Biol* 1996, **260:**289-298.
11. Schlegel S, Rujas E, Ytterberg AJ, Zubarev RA, Luirink J, de Gier JW: **Optimizing heterologous protein production in the periplasm of E. coli by regulating gene expression levels.** *Microb Cell Fact* 2013, **12:**24.
12. Mergulhao FJ, Summers DK, Monteiro GA: **Recombinant protein secretion in Escherichia coli.** *Biotechnol Adv* 2005, **23:**177-202.
13. Pritchard M, Ossetian R, Li D, Henderson C, Burchell B, Wolf C, Friedberg T: **A general strategy for the expression of recombinant human cytochrome P450s in Escherichia coli using bacterial signal peptides: expression of CYP3A4, CYP2A6, and CYP2E1.** *Archives of biochemistry and biophysics* 1997, **345:**342-354.
14. Sibbesen O, Koch B, Halkier BA, Moller BL: **Cytochrome P-450TYR is a multifunctional heme-thiolate enzyme catalyzing the conversion of L-tyrosine to p-hydroxyphenylacetaldehyde oxime in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench.** *J Biol Chem* 1995, **270:**3506-3511.
15. Smith BD, Sanders JL, Porubsky PR, Lushington GH, Stout CD, Scott EE: **Structure of the human lung cytochrome P450 2A13.** *J Biol Chem* 2007, **282:**17306-17313.
16. Barnes HJ, Arlotto MP, Waterman MR: **Expression and enzymatic activity of recombinant cytochrome P450 17 alpha-hydroxylase in Escherichia coli.** *Proc Natl Acad Sci U S A* 1991, **88:**5597-5601.
17. Bak S, Kahn RA, Nielsen HL, Moller BL, Halkier BA: **Cloning of three A-type cytochromes P450, CYP71E1, CYP98, and CYP99 from Sorghum bicolor (L.) Moench by a PCR approach and identification by expression in Escherichia coli of CYP71E1 as a multifunctional cytochrome P450 in the biosynthesis of the cyanogenic glucoside dhurrin.** *Plant Mol Biol* 1998, **36:**393-405.
18. Nielsen JS, Moller BL: **Cloning and expression of cytochrome P450 enzymes catalyzing the conversion of tyrosine to p-hydroxyphenylacetaldoxime in the biosynthesis of cyanogenic glucosides in Triglochin maritima.** *Plant Physiol* 2000, **122:**1311-1321.
19. Norholm MH, Toddo S, Virkki MT, Light S, von Heijne G, Daley DO: **Improved production of membrane proteins in Escherichia coli by selective codon substitutions.** *FEBS Lett* 2013, **587:**2352-2358.
20. Daley DO, Rapp M, Granseth E, Melen K, Drew D, von Heijne G: **Global topology analysis of the Escherichia coli inner membrane proteome.** *Science* 2005, **308:**1321-1323.

21. Lee J, Velmurugan N, Jeong K: **Novel strategy for production of aggregation-prone proteins and lytic enzymes in Escherichia coli based on an anchored periplasmic expression system.** *Journal of bioscience and bioengineering* 2013.

22. Sletta H, Tondervik A, Hakvag S, Aune TE, Nedal A, Aune R, Evensen G, Valla S, Ellingsen TE, Brautaset T: **The presence of N-terminal secretion signal sequences leads to strong stimulation of the total expression levels of three tested medically important proteins during high-cell-density cultivations of Escherichia coli.** *Appl Environ Microbiol* 2007, **73:**906-912.

23. Gotzke H, Muheim C, Altelaar AF, Heck AJ, Maddalo G, Daley DO: **Identification of putative substrates for the periplasmic chaperone YfgM in Escherichia coli using quantitative proteomics.** *Mol Cell Proteomics* 2015, **14:**216-226.

24. Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-sequence determinants of gene expression in Escherichia coli.** *Science* 2009, **324:**255-258.

25. Kahn RA, Fahrendorf T, Halkier BA, Moller BL: **Substrate specificity of the cytochrome P450 enzymes CYP79A1 and CYP71E1 involved in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench.** *Arch Biochem Biophys* 1999, **363:**9-18.

26. Laursen T, Jensen K, Moller BL: **Conformational changes of the NADPH-dependent cytochrome P450 reductase in the course of electron transfer to cytochromes P450.** *Biochim Biophys Acta* 2011, **1814:**132-138.

27. Chauhan R, Jones R, Wilkinson P, Pauchet Y, Ffrench-Constant RH: **Cytochrome P450-encoding genes from the Heliconius genome as candidates for cyanogenesis.** *Insect Mol Biol* 2013, **22:**532-540.

28. Chang MC, Keasling JD: **Production of isoprenoid pharmaceuticals by engineered microbes.** *Nat Chem Biol* 2006, **2:**674-681.

29. Klepsch MM, Persson JO, de Gier JW: **Consequences of the overexpression of a eukaryotic membrane protein, the human KDEL receptor, in Escherichia coli.** *J Mol Biol* 2011, **407:**532-542.

30. Menzella HG: **Comparison of two codon optimization strategies to enhance recombinant protein production in Escherichia coli.** *Microb Cell Fact* 2011, **10:**15.

31. Hofer R, Dong L, Andre F, Ginglinger JF, Lugan R, Gavira C, Grec S, Lang G, Memelink J, Van der Krol S, et al: **Geraniol hydroxylase and hydroxygeraniol oxidase activities of the CYP76 family of cytochrome P450 enzymes and potential for engineering the early steps of the (seco)iridoid pathway.** *Metab Eng* 2013, **20:**221-232.

32. Jorgensen L, McKerrall SJ, Kuttruff CA, Ungeheuer F, Felding J, Baran PS: **14-step synthesis of (+)-ingenol from (+)-3-carene.** *Science* 2013, **341:**878-882.

33. Mirzadeh K, Martinez V, Toddo S, Guntur S, Herrgard MJ, Elofsson A, Norholm MH, Daley DO: **Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector-Coding Sequence Junction.** *ACS Synth Biol* 2015, **4:**959-965.
34. Cavaleiro AM, Kim SH, Seppala S, Nielsen MT, Norholm MH: **Accurate DNA Assembly and Genome Engineering with Optimized Uracil Excision Cloning.** *ACS Synth Biol* 2015, **4:**1042-1046.
35. Drew DE, von Heijne G, Nordlund P, de Gier JW: **Green fluorescent protein as an indicator to monitor membrane protein overexpression in Escherichia coli.** *FEBS Lett* 2001, **507:**220-224.
36. Tsirigos KD, Peters C, Shu N, Kall L, Elofsson A: **The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides.** *Nucleic Acids Res* 2015, **43:**W401-407.
37. Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, Lerch-Bader M, Nilsson I, White SH, von Heijne G: **Molecular code for transmembrane-helix recognition by the Sec61 translocon.** *Nature* 2007, **450:**1026-1030.
38. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions.** *Nat Methods* 2011, **8:**785-786.
39. Durante-Rodriguez G, de Lorenzo V, Martinez-Garcia E: **The Standard European Vector Architecture (SEVA) plasmid toolkit.** *Methods Mol Biol* 2014, **1149:**469-478.
40. Mirzadeh K, Martinez V, Toddo S, Guntur S, Herrgard MJ, Elofsson A, Norholm MH, Daley DO: **Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector-Coding Sequence Junction.** *ACS Synth Biol* 2015.
41. Sarkar D, Le Meur N, Gentleman R: **Using flowViz to visualize flow cytometry data.** *Bioinformatics* 2008, **24:**878-879.
42. Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D, Spidlen J, Strain E, Gentleman R: **flowCore: a Bioconductor package for high throughput flow cytometry.** *BMC Bioinformatics* 2009, **10:**106.

CHAPTER 3

New cytochrome P450 homology modelling strategy identifies key amino acid residues in the CYP79A1 catalyzed conversion of L-tyrosine to (E)-p-hydroxyphenylacetaldoxime

New cytochrome P450 homology modelling strategy identifies key amino acid residues in the CYP79A1 catalyzed conversion of L-tyrosine to (E)-p-hydroxyphenylacetaldoxime

Dario Vazquez-Albacete*a, Marco Montefiori*b, Stefan Kola, Mohammed Saddikc,d, Birger Lindberg Møllerc,d, Lars Olsenb and Morten H. H. Nørholma,d#

Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Denmark a; Faculty of Health and Medical Sciences, Department of Drug Design and Pharmacology, University of Copenhagen, Denmark b, Plant Biochemistry Laboratory, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C, Copenhagen, Denmark c; Center for Synthetic Biology bioSYNergy, University of Copenhagen, Thorvaldsensvej 40, Frederiksberg C, Copenhagen, Denmarkd.

Running Title: Homology model reveals key amino acids of the CYP79A1

#Address correspondence to Morten H. H. Nørholm, morno@biosustain.dtu.dk.

Dario V.A. and Marco M. contributed equally to this work

ABSTRACT

The vast diversity and membrane-bound nature of plant P450s makes it challenging to study the structural characteristics of this class of enzymes especially with respect to accurate intermolecular enzyme-substrate interactions. To address this problem we here provide a new homology modelling strategy for structural elucidation of plant P450s guided by conserved motifs in the protein sequence and secondary structure predictions. We modelled the well-studied Sorghum bicolor cytochrome P450 CYP79A1 catalyzing the first step in the biosynthesis of the cyanogenic glucoside dhurrin. Docking experiments identified key regions of the active site involved in binding of the substrate and facilitating catalysis. Arginine 152 and threonine 534 were identified as key residues interacting with the substrate and compared with other CYP79 family enzymes by multiple alignments. The model was validated experimentally using site-directed mutagenesis and functional characterization of the heterologously expressed CYP79A1 variants in Escherichia coli. The new model provides detailed insights into the mechanism of the initial steps of cyanogenic glycoside biosynthesis. The approach could guide functional characterization of other membrane-bound P450s and provide engineering guidelines for exploitation of P450s in cell factories.

INTRODUCTION

Cytochrome P450s constitute a remarkably diverse family of membrane-associated monooxygenases that play a critical role in plant primary and specialized metabolism. In plant specialized metabolism, P450s catalyze a range of key reactions in the biosynthesis of terpenoids, alkaloids, cyanogenic glycosides and glucosinolates and phenylpropanoids. These specialized metabolites are partly formed to fend off attack from herbivores and microbial pests (1). Some of the very same compounds may therefore also be used to combat microbial infections in humans. This makes P450s highly important for biobased production of medicinal compounds because the P450s most often catalyze downstream tailoring reactions that are difficult to carry out using organic chemistry but crucial for the ultimate biological activity (2, 3). The large diversity of the P450 enzyme family in plants is observed in many ways. For example, 244 different genes encode P450s in Arabidopsis thaliana alone compared to 56 in the human genome. Curiously, P450s that catalyze the same types of reactions can display very low sequence identity (e.g. as low as 20%, (4)) and in other cases exchange of a few amino acids change their substrate specificity. Heterologous expression and structural elucidation with the current X-ray and NMR-based technologies are major bottlenecks in the study of these membrane-bound proteins. Therefore homology modeling represents an attractive alternative approach for structural studies of P450s.

Homology modeling is based on the identity/similarity of sequences between the query and the template. Thus, the low sequence identity between plant P450s and the crystalized human P450s represent a major challenge in the construction of a model. However, the comparison of available bacterial and vertebrate P450 structures has shown that the overall folding is conserved, even if the sequence identities are as low as 20 % (5, 6). In particular, seven α-helices and four β-pleated sheets that contribute to the overall fold are well conserved in the structure but not always predictable from the primary sequence (5, 6). Other regions in the P450 enzymes, involved in the binding of specific substrates, are less conserved. Six of these regions have previously been labeled Substrate Recognition Sites 1 to 6 (SRS1-6) (6). Given the high importance of these regions, it is important to model SRS1-6 with particular attention.

Cyanogenic glucosides are examples of hydrogen cyanide-releasing chemicals that play important roles in defense against herbivore predation (7). Upon tissue disruption, caused by a chewing insect or livestock, the cyanogenic glucosides are brought in contact with endogenous β-glucosidases and hydrolyzed with concomitant release of hydrogen cyanide. The accumulation of cyanogenic glucosides in certain crops such as forage sorghum (Sorghum bicolor) is problematic when used for animal feed due to the associated toxicity of HCN (8). The understanding of the process of cyanogenesis is therefore of theoretical as well as practical importance. Oximes are key intermediates in the biosynthesis of cyanogenic glucosides (9). The oximes are synthesized from a small selected number of amino acids in a reaction involving two N-hydroxylations, a decarboxylation and a dehydration step (10, 11). In all plant species currently investigated including

monocots as well as dicots, these reactions are catalyzed by a multifunctional P450 enzyme belonging to the CYP79 family (4). This P450 family is known to have a high specificity for their substrates, mostly natural amino acids, although the mechanistic bases for this specificity still remain unclear (12-14) . The CYP79A1 catalyzes the committed step in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (15, 16). The functional properties of CYP79A1 have been studied using sorghum microsomal preparations and following incorporation in detergent micelles or nano discs, and has become a model for studying plant P450 function. Crystallization of CYP79A1 to further advance the structural studies has not been achieved. CYP79A1 has very high substrate specificity, L-tyrosine being the only substrate found, whereas the second enzyme of the dhurrin pathway, CYP71E1 accepts several different oximes as substrates (16). A combined biochemical screen and TILLING approach involving more than two thousand independent sorghum lines in the field pointed to the functional importance of specific amino acid residues in CYP79A1. Sorghum plants harboring the mutations E145K and P414L located near SRS-1 and SRS-6, respectively, produced no or reduced amounts of dhurrin (8). Attempts have been made to model how CYP79A1, CYP71E1 and the partnering reductase (POR), may interact and catalyze all the biosynthetic reactions in an enzyme complex often referred to as a metabolon (17). These computational studies suggested that the R152 residue located in the BC-loop region, a flexible region between the B-helix and the C-helix, participated in the binding of tyrosine although this was not validated experimentally.

In the current study, we outline a systematic approach to improve homology modeling of P450s using CYP79A1 as a model enzyme and validate the approach by determining the activity of the microbially expressed mutant enzymes. We also investigate the mechanistic bases of substrate binding and specificity combining site-directed mutagenesis, structural data and multiple sequence alignments.

MATERIALS AND METHODS

Homology modelling, sequence analysis and docking. The primary sequence of CYP79A1 was downloaded from uniprot (ID Q43135). The first 68 residues of the protein consist of a high variable region including the transmembrane domain (13-33 amino acids). Given the low number of P450 crystal structures that include the transmembrane domain, and the large distance from the catalytic site (Fig.1), this part of the sequence was removed prior to modeling to avoid misaligning of the templates. Given the low identity score between the query and the available 3D structure templates (the top result showed 27% sequence identity) we decided to use a modified “hybrid-structure” approach (18). The sequence was divided into three different regions, each containing conserved motifs and variable regions specifically including the Substrate Recognition Sites (SRSs). The first region (residues 60-168) contained SRS-1. The second region (residues 168-310) contained the conserved consensus WXXXR motif (part of the heme cradling architecture (19, 20),

SRS-2, SRS-3, SRS-4 and SRS-5. The third region (residues 310-558) contained SRS-6. The secondary structure for the protein was predicted using Predictprotein (21). Templates for the homology model were chosen based on two different elements: the sequence identity in alignments with the BLOSUM62 matrix and the similarity of the secondary structure. The homology model was developed with the Chimera (v. 1.9) interface for Modeller (v. 9.15) (22, 23). Ten different models were generated, and the model with the best z-dope score was selected.

Docking was performed using Glide (v. 6.3; Schrödinger)(24). Receptor grids were constructed around the iron of the heme group, with sizes of 10 x 10 x 10 Å. Grids were defined both with and without metal-binding restraints. Ligands were docked in the receptor grid using extra-precision, flexible ligand sampling mode. Tyrosine was docked in neutral charge form in the active site, which is the dominating form at physiological pH. To evaluate the best docking poses, the Glide docking score was used.

For sequence comparison, multiple alignments were carried out with experimentally characterized CYP79s found in the literature and Uniprot data base. In addition we took into account phylogeny of these protein sequences using a maximum likelihood approach as previously described (25).

PCR and uracil excision. Uracil excision cloning was performed as previously described (26). Details of oligonucleotides, template DNA and references can be found in Tables S1 and S2 and below. PCR products were amplified in 50 µL reactions containing: 1 µL PfuX7 DNA polymerase , 0.2 mM dNTPs (Thermo Scientific, Waltham, USA), 1.5 mM MgCl2, 0.5 µM forward oligonucleotide, 0.5 µM reverse oligonucleotide, Phusion® HF Reaction Buffer (New England BioLabs, Ipswich, USA) and 50 ng plasmid template. A touch-down PCR program was used for amplification: Step 1: 2 min 98°C; step 2:15 sec 98°C, 20 sec 65°C (-1°C per cycle), 45 sec per kb at 72°C (step 2 repeated 9 times until 70°C, then repeated 20 cycles at the annealing temperature 60°C); step 3: 5 min 72°C; step 4 hold at 10°C. PCR products were gel purified from 1% (w/V) agarose gel using NucleoSpin Gel and PCR Clean-up (Macherey-Nagel, Düren, Germany) and eluted in nuclease free water. Purified PCR products were incubated with 1 µL USERTM enzyme (New England BioLabs, Ipswich, USA) for 30 min at 37°C and at 20°C for 15 min. A Nanodrop spectrophotometer 2000 (Thermo Scientific, Waltham, USA) was used for estimation of PCR product and vector concentration. Approx. 5 µL of the uracil-excision reaction product solution was transformed into NEB 5-alpha chemically competent cells according to the manufacturer's protocol. Transformants were selected on Luria Bertoni (LB) agar plates supplemented with 50 □g/mL kanamycin and 10 □g/mL gentamycin. Colonies were screened for positive mutants by DNA sequencing (Eurofins Genomics, Ebersberg, Germany).

DNA constructs and site directed mutagenesis. Briefly, in order to facilitate gene overexpression and membrane targeting, a sequence encoding for the signal peptide of the endogenous *E. coli* periplasmic protein disulfide isomerase dsbA MKKIWLALAGLVLAFSASAAQ (BAE7748) was added by one-fragment uracil-excision cloning (27) to the 5´ end of CYP79A1 (U32624) in a pET28a(+)-derived construct. In this vector, in addition to the dsbA sequence, CYP79A1 is fused to sequences encoding a TEV protease site, GFP folding reporter and a polyhistidine tag (28). Similarly, site-directed mutagenesis was performed by one-fragment uracil-excision cloning with the previous construct as template to introduce alanine substitutions in the desired amino acid positions (Table 1). A Strep-HRV3C tagged, codon optimized Sorghum bicolor POR2b (Wadsäter et al., 2012) was cloned into the pET28a(+)-tev-gfp-his8 construct and gfp was subsequently deleted and the origin of replication (ori) replaced with the corresponding ori from pSEVA63 (Silva-Rocha et al., 2013) by amplifying pSEVA63 with the oligo nucleotides 5´-ATCCGCTUTAATTAAAGGCATCAAATAAAAC-3' and 5´-ACTAGTCTUGGACTCCTGTTGATAGATC-3' and the pET28-based por2b construct with the oligo nucleotides 5´-AAGCGGAUCTACGAGTTGCATGATAAAGAAGACAGTC-3' and 5´-AAGACTAGUCAATCCGGATATAGTTCCTCCTTTCAG-3' followed by uracil excision cloning.

TABLE 1 Amino acid mutations included in the present study

Mutant	Predicted function	Reference
R152A	Substrate binding	This study and (17)
E145K	Substrate recognition	(8)
D347A	Binding of intermediate	This study
D354A	Catalysis	This study
N355A	Catalysis	This study
P414L	Structural	(8)
T534A	Substrate binding	This study

Protein production and purification. All DNA constructs were transformed into the chemically competent *E. coli* KRX strain (Promega, Madison, USA [F´, traD36, ΔompP, proA+B+, lacIq , Δ(lacZ)M15] ΔompT, endA1, recA1, gyrA96 (Nalr), thi-1, hsdR17 (rK –, mK+), e14– (McrA–), relA1, supE44, Δ(lac-proAB), Δ(rhaBAD)::T7 RNA polymerase) for gene expression. Cells were grown on Terrific Broth (TB) (1.2% tryptone; 2.4% yeast extract; 0.4% glycerol; 17 mM KPi (monobasic); 72 mM KPi (dibasic)). Over-night cultures were prepared with media supplemented appropriate antibiotics. The optical density (OD) of the over-night cultures was measured at Abs600nm in a Plate Reader (SynergyMx, SMATLD) (BioTek, Winooski, USA). The pre-cultures were subsequently inoculated into 400 mL fresh TB medium in 2 L shaking flasks at a final Abs600nm of 0.05 and grown at 30°C, 250 rpm in an Innova®44R incubator shaker

system (5 cm orbital shaking) (New Brunswick Scientific, Eppendorf, USA) to an optical density of 0.5. Expression was induced by the addition of isopropyl ß-D-1-thiogalactopyranoside (IPTG, dioxane free, Thermo Scientific, Waltham, USA) at a final concentration of 0.4 mM and 5 mM L-rhamnose (Sigma-Aldrich, St. Louis, USA). Cultures were subsequently incubated at 25°C, 250 rpm for 20 h. The cells were harvested (2,500 x g, 10 min at 4˚C), washed twice in 20 mM Tris buffer pH 7.8 and resuspended in 90 ml of lysis buffer consisting of 20 mM Tris with 250 U/mL Benzonase ® nuclease (Sigma-Aldrich, St. Louis, USA) and the complete ULTRA EDTA-free protease inhibitor cocktail (Roche, Basel, Switzerland). To recover CYP79A, membranes were isolated as previously described (29). Cells were broken with a French press Emulsiflex C-50 (Avestin Europe GmbH) with minor modifications, at 15,000-20,000 p.s.i. for at least two passes at 4°C. Unbroken cells were removed by centrifugation (20,000 x g, 20 min at 4˚C) and the supernatant collected for further fractionation. The membrane fraction was separated by centrifugation (150.000 x g, 45 min at 4°C), membranes were recovered from the pellet in 20mM Tris buffer pH 7.8, 20% glycerol and quickly frozen at -80˚C. For protein purification, membranes were solubilized with 10% DDM (Sigma-Aldrich, St. Louis, USA) for 4 h at 4˚C. Unsolubilized material was removed by centrifugation at 20,000g for 45 min at 4˚C and the supernatant applied onto nickel-nitrilotriacetic acid (Ni2+-NTA) resin columns (HisTRAP) on an Äkta Pure system connected to an F9-C fraction collector (General Electric, New York, USA). The bound fusion protein was washed extensively with IMAC buffer containing 0.1% DDM (w/v) and was subsequently eluted by increasing the imidazole concentrations to 500 mM in a single step. The fractions containing the protein of interest were combined and concentrated to 0.5 mL. Next, gel filtration was used to separate the fusion protein from cleaved-off GFP. The sample was injected onto a Superdex 200 Increase 10/300 GL column equilibrated with 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10% glycerol, 0.1% DDM. The fractions containing the fusion protein were combined, analyzed by SDS-PAGE, flash-frozen in liquid nitrogen and stored at −80°C for subsequent analysis.

Whole cell fluorescence measurements. Whole cell fluorescence was measured using 2 mL induced culture. The cells were harvested (2,500 x g, 20 min) and resuspended in a total of 200 □L PBS buffer. Fluorescence was detected using excitation at 485 nm and emission at 512 nm with a window of +/- 9 nm, using a plate reader (SynergyMx SMATLD, BioTek, Winooski, USA). Expression levels were estimated with a purified GFP standard diluted in cells with the empty vector as previously reported (30).

Protein analysis. Protein samples were analyzed with PageRuler™ Prestained Protein Ladder 10-170K (Thermo Scientific, Waltham, USA) by standard SDS-page using Mini-PROTEAN® TGXTM 4-15% gels (Bio-Rad, Hercules, USA).

Enzyme activity measurements. For the in vivo analysis of CYP79A1 activity, cells were harvested by centrifugation at 2,500g, 4°C for 10 min, washed once in 50 mM KPi pH 7.5, adjusted to the same to OD and used in 50 mM KPi. The enzyme reaction was carried out in 200 µL volume consisting of 5 mM NADPH, 0.1 mM L-Tyrosine, Phenylalanine or N-methyl-tyrosine (Sigma-Aldrich, St. Louis, USA), 50 mM KPi and 180 µL cell suspension. Cells were incubated at 30°C, 250 rpm for 60 min. The product (E)-p-hydroxyphenylacetaldoxime (oxime) was extracted with 1 volume of methanol. Cells were removed by centrifugation (20,000 x g, 10 min) and the supernatant was transferred into HPLC vials and stored at -20°C prior to LC-MS analysis.

In vitro activity assays were carried out in 50 µL 5 mM NADPH, 0.1 mM L-Tyrosine (Sigma-Aldrich, St. Louis, USA), 50 mM KPi, 0.3 mg/ml of purified CYP79A1 and 0.3 mg/ml of purified PORb2 kindly provided by Dr. Tomas Laursen (University of Copenhagen, Department of Plant and Environmental Sciences) POR2b (31). The reaction was incubated at 30°C at 250 rpm for 60 min. The oxime was extracted with 50 µL of methanol and samples stored at -20 prior to analysis by LC-MS.

LC-MS detection. The oxime (MW 151.17) was detected and quantified by LC-MS. A chemically synthesized standard was kindly provided by Mohammed Saddik Motawie (University of Copenhagen, Department of Plant and Environmental Sciences) (32).

The (E)-and (Z)-oximes (the two geometrical isomers) were detected in both the standard and the assay due to instability and chemical equilibrium (9, 33).Thus for the purpose of total oxime quantification the chromatogram area of both peaks were summed. LC-MS data was collected on a Bruker Evoq triple quadropole mass spectrometer equipped with an Advance UHPLC pump system. Samples were held in the CTC PAL autosampler at a temperature of 10°C during analysis. Injections (2 µL) of the sample were made onto a Discovery HS F5-3 HPLC column (3 µm particle size, 2.1 mm i.d., 150 mm long). The column was held at a temperature of 30.0 °C. The solvent system (flow rate: 1.0 ml/min) used was water with (A) 100mM ammonium formate and (B) acetonitrile using the following elution profile: 0.5 min 95% A/5% B, linear gradient to 50% A/50% B for 3.0 min, 1.5 min 50% A/50% B and re-equilibration for 2 min 95% A/5% B. The column eluent flowed directly into the heated ESI probe of the MS, which was held at 350°C and a voltage of 4500 V. SRM data was collected in centroid at unit mass resolution. Positive ion mode with Q1 set to monitor 152.70 m/z, Q3 set to monitor 136 m/z and Q2 set to a collision energy of 10.0eV, with an Argon pressure of 1.5 mTorr. The other MS settings were as follows, Sheath Gas Flow Rate of 40 units, Aux Gas Flow Rate of 40 units, Sweep Gas Flow Rate of 20 units, Ion Transfer Tube Temp was 350 °C.

RESULTS

Homology Modeling. For the modeling, the primary sequence of the enzyme was divided into three different modules. For each separate module, the most similar template that met the dual requirement of high primary sequence and predicted secondary sequence resemblance was chosen. The templates found to fulfill both criteria for the different regions in the Protein Data Bank were: 1SUO (31% identity) for the first section, 3MZS (21% identity) and 3CZH (18%) for the second region, 2HI4 (35% identity) for the third region. An additional template was chosen to model the overall protein structure was 4I8V (27% identity). SRSs are usually flanked by structurally conserved regions, and these were used to guide the homology modelling, with the aim of improving accuracy of the overall structure prediction. The six SRSs were located with reference to the specific conserved secondary sequence regions. By prediction of the secondary conserved regions (SCRs), the position of the SRSs was identified in the primary sequence. This facilitated the alignment and the choice of the templates used to build the homology model. The main conserved structural features, the heme-binding domain (WXXXR), as well as I helix that flank the substrate cavity were identified. SRS-1 was located in the loop between the conserved B and C helices (called BC-loop) while SRS-6 was located in the proximity of the carboxyl terminal of the protein, flanked by two conserved regions, referred to as SCR-17 and SCR-18 (34). The 6 SRS regions consist mainly of loops, except SRS-4, which is located inside one of the conserved helices. The homology model obtained is shown in the Fig.1.

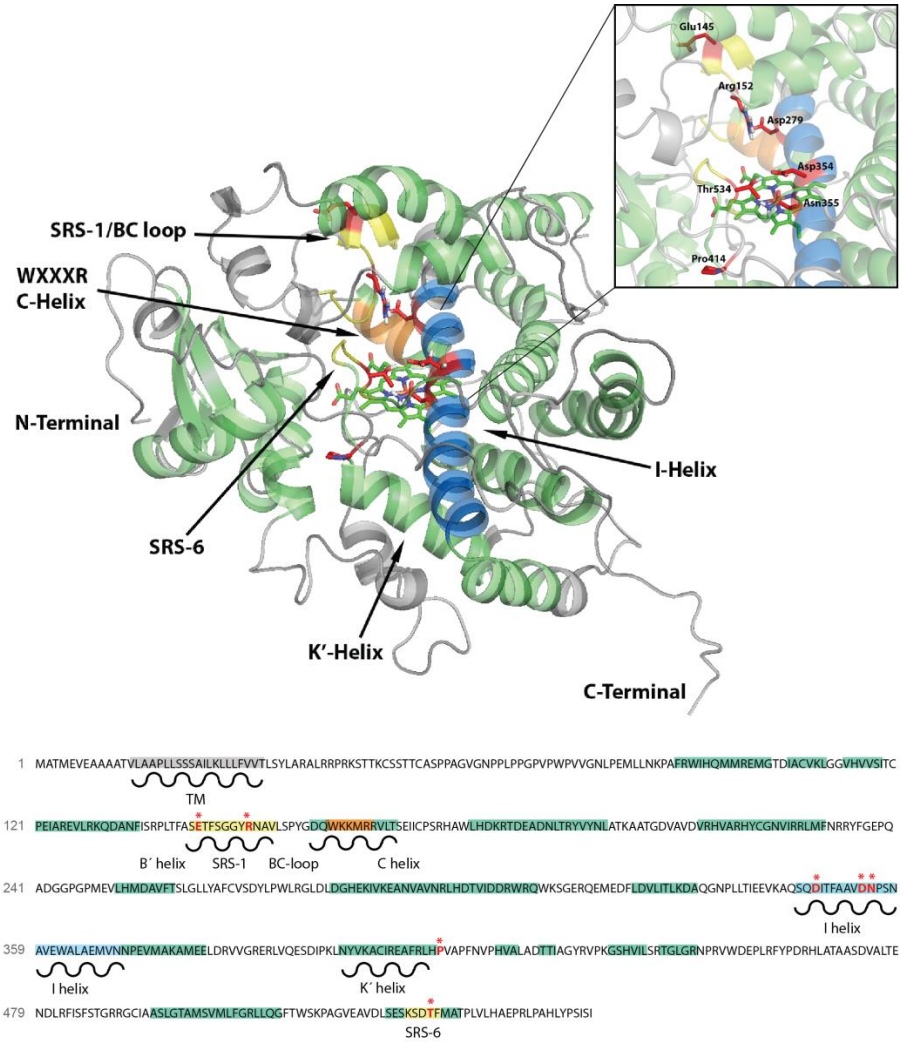


FIG 1 Homology model of *Sb*CYP79A1 with regions under study highlighted in the primary sequence. Grey: non-conserved regions, green: conserved regions, orange: WXXXR consensus motif, yellow: Substrate Recognition Site SRS-1 and SRS-6, red: residues mutated in the study. Top-right square with a zoomed view of the mutated residues in the study.

Docking. A docking study was carried out using the natural substrate of the enzyme in order to identify possible regions involved in binding and catalysis. Two hydrogen bonding interactions were observed. Docking results show that R152 is located close to the BC-loop and whose positively charged nitrogenous guanidinium group interacts with the carboxylate group of the tyrosine substrate (Fig. 2A). This interaction serves to stabilize the position of the carboxyl group and thus the docking position of tyrosine in the active site. The region where this Arg-mediated contact takes place is close to the substrate recognition site (SRS-1) as suggested in a previous model (17). The second interaction involves T534 located in the surroundings of the predicted SRS-6. This interaction stabilizes the para-hydroxy group of tyrosine by hydrogen bond distance from 2.6 Å. The positioning of the para-hydroxy group of tyrosine is further stabilized by hydrogen

bonding to N355. The first step in the CYP79A1 catalyzed reactions is conversion of L-tyrosine into N-hydroxytyrosine by hydroxylation of the amino group. Docking of N-hydroxytyrosine, revealed the same two residues, R152 and T534 (Fig 2.B). In addition a new transient interaction between D347 situated in the I-helix and the incorporated N-hydroxyl group of the N-hydroxytyrosine is formed (Fig. 2.B). As a result of this interaction the rotation around the C-N bond of the tyrosine amino group is prevented as previously suggested (35). Finally, we identified two potentially relevant amino acids involved in the oxygen activation cycle of P450s positioned within the I-helix in the proximity of the para-hydroxyl group of tyrosine (Fig.1): D354 and N355. The catalytic cycle of a P450 involves substrate binding and stabilization in the active site followed by binding of molecular oxygen coordinated to the iron atom of the heme group (36). The bound oxygen undergoes activation through nucleophilic dioxygen intermediates. In this process, neighboring amino acids such as aspartate or asparagine can assist several protonation and deprotonation steps of dioxygen intermediates (37). Finally, the highly reactive iron-oxo electrophilic species catalyzes hydroxylation of the substrate. This cycle provides the mechanism for substrate hydroxylation and thus has to be accounted for in structural models. Multiple alignments to other CYP79 revealed the strict conservation of these two residues (Fig. 3A). In the model, the amino group of N355 is located at a distance of 2.9 Å from the para-hydroxyl group of tyrosine. Finally, D354 and N355 present proton acceptor and donor capabilities respectively, an essential feature in the activation of molecular oxygen (36). In addition to the aforementioned residues, the latter residues were also selected for experimental validation of the model (Table 1). The distance between the iron atom in the heme group and the amino group of the tyrosine is 4.0 Å, the expected distance for N-hydroxylation to happen (38, 39).

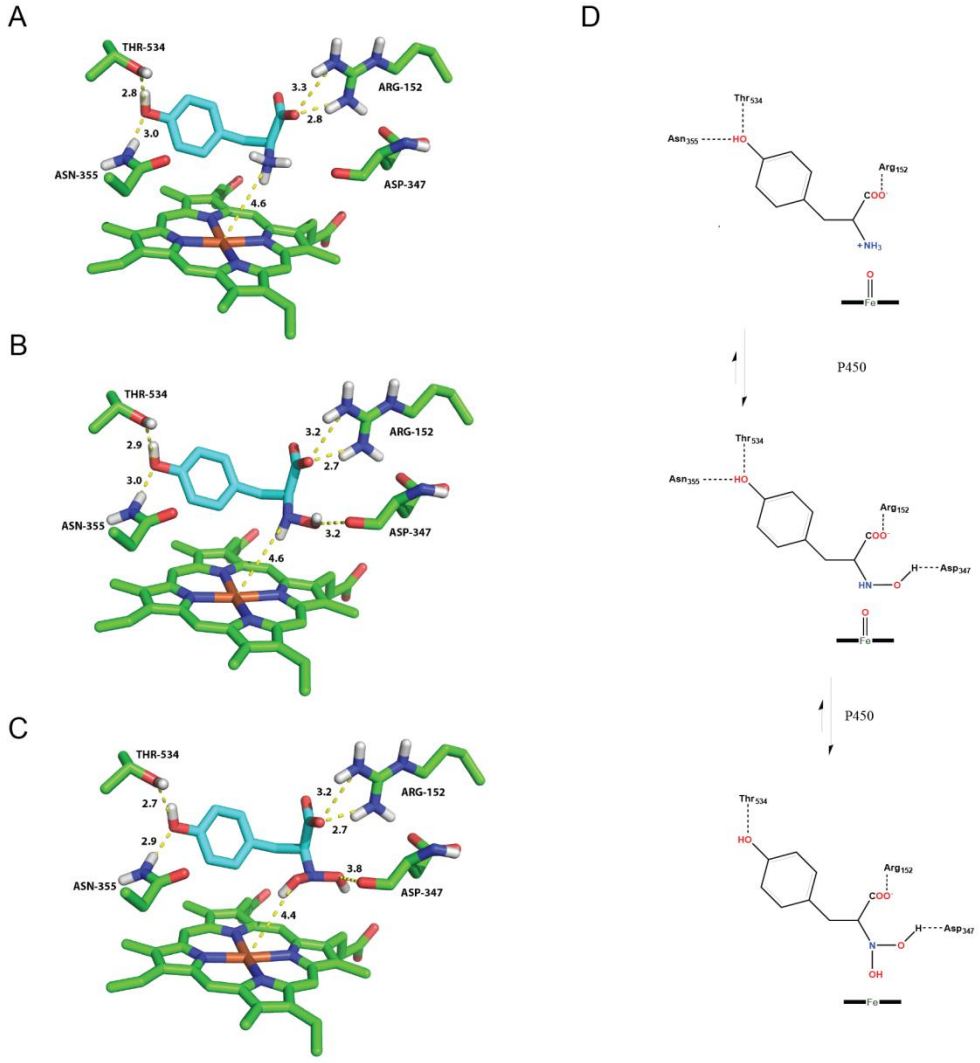


FIG 2 Illustration of the interactions between SbCYP79A1 and tyrosine identified by docking. The unprotonated carboxyl group of tyrosine is shown within hydrogen bond distance from R152. The para-hydroxyl group of tyrosine appears stabilized by T534. The Catalytic amino acid N355 is also observed close to the hydroxyl group of tyrosine (A). For the first intermediate of the reaction, N-hydroxytyrosine, the same two interactions are shown together with the newly formed interaction between the incorporated hydroxyl and D347 (B). Docking pose of the last intermediate of the reaction catalyzed by the CYP79A1, N´N-dihydroxytyrosine stabilized by R152 and T534 is shown. The interaction between the hydroxyl group and D347 is also observed (C). Chemical representation of the stabilizing contacts between the CYP79A1 and tyrosine followed by the subsequent hydroxylation steps catalyzed by the enzyme (D).

Two additional substrates were docked in the CYP79A1 to validate the interactions identified and understand the mechanisms of enzyme specificity. For this, a commercially available tyrosine analogue called N-methyl-methyltyrosine was docked in the active site using the same settings. This analogue is decorated with a methyl group branching out from the amine group of tyrosine. Interestingly docking of this analogue yielded

the same interactions as for tyrosine (Fig S1). The three amino acid residues R152, T534 and N355 appear to preserved. However, no contacts were observed with the D347. In the present study phenylalanine was also docked in the active site. Previous studies demonstrated that this amino acid is not converted to oximes by the CYP79A1 (16). The docking results show that only a weak interaction with N355 is established with phenylalanine (Fig S1) (39)(14).

Protein production and purification. Plant P450s are considered difficult-to-express in *E. coli* possibly because their hydrophobic N-termini are destined for insertion into the plant endoplasmic reticulum and therefore may not be recognized by the bacterial cytoplasmic membrane translocation machinery. The wild type CYP79A1 sequence is not expressed in *E. coli* as happens with other P450s from this family (40). Thus, in this study we created a CN-terminal fusion with a small bacterial signal sequence from the DsbA protein (DsbAss), which has previously been successfully used for similar purposes (41). In addition, a special GFP fluorochrome that enables for reporting properly folded membrane proteins facing the cytoplasmic side, was fused to the c-terminal (28). Unlike the heterologous P450 sequence, the DsbA signal sequence should be recognized by the bacterial translocation machinery and targeted to the periplasm using the signal recognition particle (SRP) pathway in *E. coli.* When the DsbAss is translocated across the membrane, the native transmembrane region of CYP79A1 will likely be anchored in the inner membrane of *E. coli* leaving the catalytic domain at the cytoplasmic side of the cell and positive GFP fluorescence. Large expression cultures were set up for all CYP79A1 mutants all being expressed as DsbAss fusions. Bacterial membranes were prepared from whole lysates, detergent solubilized and the mutated CYP79A1 proteins purified to homogeneity using His-tag affinity columns and gel filtration. Although the DsbA signal peptide provides a solution to solve the functional expression bottleneck, it renders a 420nm CO spectra (data not shown), making the GFP fusion an alternative and more accurate method to estimate enzyme concentration.

Enzyme assays. The new homology model of CYP79A1 suggested that R152 and T534 play roles in substrate and intermediate binding, whereas D354 and N355 may be involved in catalysis. Previously, mutations E145K and P414L were also suggested to affect the activity of the enzyme based on measurements of cyanide release from leaves of sorghum TILLING mutants (8). In vivo assays were performed in intact *E. coli* cells co-expressing each of the genes encoding the six mutant CYP79A1 enzymes and the natural electron donor partner POR2b (31). All mutants were expressed as GFP fusions with GFP and fluorescence was used to normalize expression levels because DsbA fusions produce negative 450nm CO spectra. Three out of the six mutants were active (Table 2). Compared to the wildtype control, mutants R152A and E145K located in the BC-loop and SRS-1 region displayed 35 and 48% residual activity compared to wildtype CYP79A1, respectively. The two mutant CYP79A1 enzymes with the D354A or N355A substitution in the I-helix identified as catalytic residues showed no detectable hydroxylation activity. The mutant T534A showed 61% residual activity compared to wildtype CYP79A1. The P414L

mutant previously demonstrated to cause homozygous sorghum seedlings to be acyanogenic had no residual activity (8). The partially active CYP79A1 mutant proteins were chosen for protein purification and further characterization in vitro. The isolated proteins were analyzed by SDS-PAGE to verify the correct size of the fusion protein (Fig S2).

TABLE 2 *In vivo* activity of *Sb*CYP79A1 mutants and wild type control in intact *E. coli* cells analysed by LC-MS.

Mutant	Activity[a] µU/ml	% residual activity
R152A	36±22	35%
E145K	48 ±3	48%
D354A	*ND*[b]	0%
N355A	*ND*[b]	0%
P414L	*ND*[b]	0%
T534A	63 ±9	61%
Control	103 ±20	100%

[a]*Enzyme activity in IU*

[b]*ND*: No activity detected

The isolated pure CYP79A1 mutant proteins were incubated with purified PORb2, NADPH and 0.1mM tyrosine. The D347A mutant was also purified and compared in vitro with the other active mutants. In line with the results obtained using in vivo assays, reduced aldoxime production was observed for the R152A and E145K mutant CYP9A1 compared to the wildtype control (0.1%), whereas the T534A and D347A mutants exhibited 15.1% and 9.6% respectively of wildtype activity (Table 3).

TABLE 3 *In vitro* activity of purified *Sb*CYP79A1 mutants compared to the wild type control analysed by LC-MS.

Mutant	Activity[a] µU/ml	% residual activity
R152A	2 ±1	0.1%
E145K	2±1	0.1%
D347A	34±10	9.6%
T534A	53±16	15.3%
Control	354±150	100%

[a]*Enzyme activity in IU*

Although the purification process reduces enzyme activity compared to in vivo experiments, the trend is maintained. The GFP reporter platform indicates that all mutants are expressing to similar levels around 100mg/L in the concentrated broth except for the N355A mutant, which displayed reduced expression (Fig.S3 and S4). To investigate the involvement of the D347 and T534 residues in substrate specificity, phenylalanine and N-methyltyrosine were tested as alternative substrates in the wild type and T534A mutant respectively. None of these substrates were converted to oximes by neither the wild type nor the T534A mutant (Table 4).

TABLE 4 *In vitro* activity of purified *Sb*CYP79A1 mutants compared to the wild type control with different substrates analysed by LC-MS.

Mutant	Substrate	Activity[a] µU/ml
T534A	Phenylalanine	*ND*[b]
Control	Phenylalanine	*ND*[b]
Control	N-methyltyrosine	*ND*[b]

[a]*Enzyme activity in IU*

[b]*ND*: No activity detected

Comparative analysis of the CYP79 family. To elucidate structural features involved in conservation or evolution of the substrate binding mechanism of the CYP79A1, we integrated homology model data, experimental results, primary sequence alignments and philogeny with other experimentally characterized CYP79 enzymes. The following enzymes met these criteria; TmCYP79E1, TmCYP79E2, AtCYP79F1, AtCYP79F2, SaCYP79B1, AtCYP79B2, AtCYP79B3, MeCYP79D2, MeCYP79D1, AtCYP79A2, LjCYP79D3 and LjCYP79D4.

All CYP79 sequences have between 51% and 59% identity with the wild type CYP79A1. Importantly, several observations are made from alignments of the regions under study i) The closest CYP79 family group to the CYP79A1 is the CYP79B from Arabidopsis thaliana and Sinapis alba. This group has narrow specificity for tryptophan (11, 40), a lysine is found instead of arginine (same chemical properties) compared to the 152 position of the CYP79A1, glutamate instead of aspartate, also with the same chemical properties in the 347 position and methionine replacing threonine (Fig. 3A) ii) Interestingly cytochromes CYP79E1 and CYP79E2 are the only two characterized enzymes where arginine and aspartate are aligned in the same position as in the CYP79A1. These two cytochromes convert tyrosine, the same substrate of the CYP79A1, to oximes in a very specific manner in Triglochin maritima (14). iii) All characterized CYP79 enzymes have a conserved aspartate-arginine pair in the predicted I-helix, identified as responsible for the oxygen

activation cycle. iv) Finally, the CYP79A1 exhibits a unique threonine amino acid residue in the position 534 compared to the rest of aligned sequences.

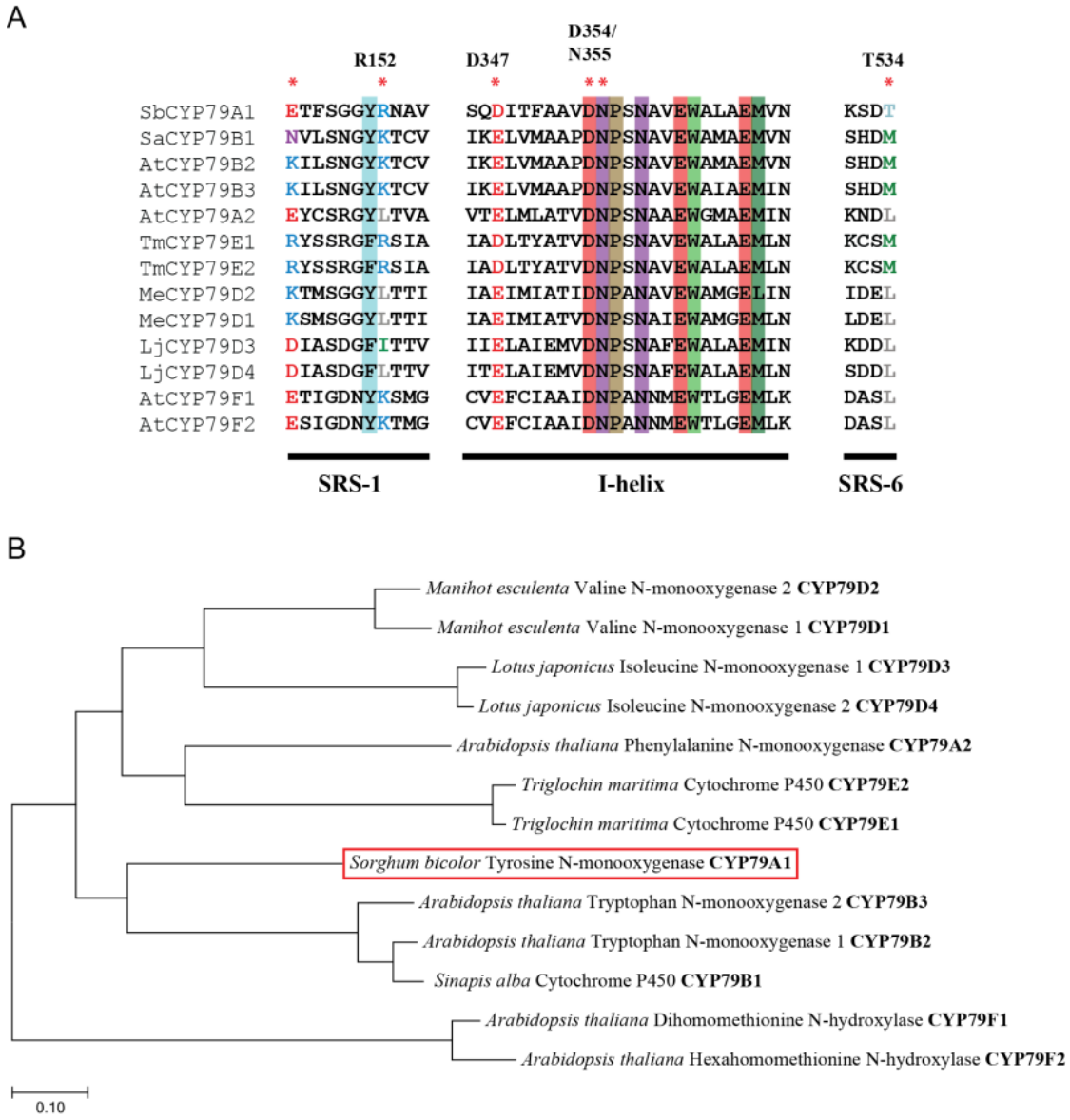


FIG 3. Sequence comparison of the CYP79 family. Multiple alignment of the important regions identify in the docking (A). Phylogeny tree of characterized CYP79 enzymes by maximum likelihood (B). Sb: *Sorghum bicolor*, Sa: *Sinapis alba*, At: *Arabidopsis thaliana*, Tm: *Triglochin maritima*, Me: *Manihot esculenta*, Lj: *Lotus japonicus*

DISCUSSION

Due to the importance of P450s in the biosynthesis of a number of desired high value plant natural products, it has become indispensable to understand what structural features are responsible for a given function. In this sense homology modeling is an important tool to better understand the relations between structure and activity of an enzyme. However, when the amino acid sequence identity is low (below 30%) it is difficult to determine whether a structural model obtained is trustworthy or not. In this case, other approaches have to be pursued to build a better model. Here, we describe a new method to increase the accuracy of homology models for P450s to possibly accelerate functional characterization of members of this enzyme class. To test the validity of the approach, we studied CYP79A1 as a model plant P450. Mutations and amino acids to which a function was assigned in previous work were included for benchmarking and for completeness of the analyses. Our new homology model of CYP79A1 was used to characterize the active site by docking techniques. We determined amino acids involved in binding and catalysis of the tyrosine substrate and the catalytic residues mediating its N-hydroxylation. Likewise the mechanism of N-hydroxylation carried out by the CYP79A1 was reconstructed. Furthermore, we validated the model by measuring mutant enzyme activity combined with a GFP reporter platform to monitor the obtained CYP79A1 protein expression levels in E. coli.

The approach for the homology modeling of the P450 relies on a combination of different techniques. The model was built using a multiple templates approachThis approach had generally been used in cases in which a single template failed to provide complete structural information of the target (42). Given the low sequence identity among the available P450 crystal structures, it was decided to use the multi-template approach to possibly build a more reliable model of the CYP79A1 enzyme and to identify the key residues in the enzyme-substrate interaction. Four different templates where used: a general template, for modeling the entire protein structure, and different templates to model three different regions of the enzyme. In choosing the templates, the sequence identity to the target as well as the correspondence between the actual secondary structures the template and the predicted secondary structure of the target was considered. The integration of the predicted and known structural data in the model like the secondary structure, consensus motifs and SRSs guided the building of an accurate model. The use of substrate docking to identify key interactions between the enzyme and the substrate is strongly dependent on the quality of the modelled structure. The good correlation between the docking studies conducted using our homology model and the experimental studies on E. coli expressed wildtype and mutated CYP79A1 validated the robustness of our approach. The detailed structural data obtained enabled identification of the most important peptide regions and amino acid residues that govern the interaction between the enzyme and its specific substrate. Most importantly, the robustness of our model helped to decipher the possible tyrosine´s hydroxylation mechanism catalyzed by the CYP79A1 in agreement with previous biochemical evidence based on in vitro studies using sorghum microsomes (43).

These studies showed that the conversion of tyrosine to p-hydroxyphenylacetaldoxime was accompanied by the consumption of two molecules of 18O2. T In addition, incubation in an 18O2 atmosphere in the presence of tyrosine or N-hydroxytyrosine as substrate demonstrated that the 18O atom incorporated in the second N-hydroxylation reaction using N-hydroxytyrosine as substrate was completely retained in the subsequent production of the p-hydroxyphenylacetaldoxime. This demonstrate that all intermediates in the CYP79A1 catalyzed conversion are tightly bound within the catalytic site of CYP79A1 permitting no rotation around the C-N bound of the intermediates (35). These biochemical observations may now be rationalized based on the CYP79A1 protein model.

The new CYP79A1 model suggests three different specific interactions between the tyrosine substrate and the CYP79A1 enzyme. First the positively charged guanidinium function amino group of R152 interacts with the negatively charged carboxyl group of tyrosine. R152 is located in the BC-loop harbored within SRS-1. Several studies have demonstrated the role of this whole region in determining the stereospecificity of P450s (34) and (44). In the present study we demonstrate experimentally that an R152A substitution dramatically reduced enzyme activity in agreement with previous models (Table 2 and 3). The interaction of R152 with the carboxylate group of tyrosine prevents its free movement, thus may provide stability for the first hydroxylation round (Fig. 2A). The hydroxylation of the guanidinum nitrogenous group is hypothesized to occur through a hydrogen transfer and rebound mechanism, which takes place at a molecular distance of ~5Å in agreement with the docking results (39). Despite homology of only 53% ,alignment of CYP79 sequences revealed that only CYP79E1 and CYP79E2 from Triglochin maritima have this position conserved in comparison with the CYP79A1 (Fig. 3A). Because these two enzymes specifically catalyze the conversion of tyrosine to oximes with the same degree of preference for the substrate, it is possible that this region may be indeed involved in the specific binding (14). In addition, the CYP79B1, CYP79B2 and CYP79B3, all with remarkable substrate specificity (40), display the amino acid lysine with the same chemical properties in this position. Thus, similar stabilization mechanism of the carboxylate group may occur in these enzymes. The second interaction discovered in the docking involves T534, which forms a hydrogen bond to the para-hydroxy group of tyrosine and is located very close to SRS-6. The T534A mutant exhibited lower activity compared to the wildtype, but not as low as with mutations in the SRS-1 and BC-loop (Table 2 and 3). This region has previously been suggested to be involved in binding of substrates in other eukaryotic P450s (45). In CYP79A1, stabilization of the positioning of the aromatic tyrosine ring did not seem critical judged by the minor effect of introducing the T543A mutation. This may be explained by the tight docking taking place based on the positioning of the para-hydroxy and carboxylate groups (Table 2 and 3) (17). Curiously among all CYP79 with known activity, the CYP79A1 is the only one with a threonine residue in this position. The enzymes CYP79E1, CYP79E2 and CYP79B group exhibit methionine in this position. All these enzymes catalyze the conversion of aromatic amino acids to oximes. However, the CYP79A2 from Arabidopsis thaliana also catalyzes the conversion of the aromatic amino acid phenylalanine to its respective derived

oxime but displays a leucine not only in the T534 position but also in the R152 (Fig 3A) (46). One would expect that replacement of a polar amino acid like T534 by a non-polar amino acid such as alanine, could replicate the environment of the CYP79A2 and turn it into a phenylalanine hydroxylating enzyme. Nevertheless no conversion to oxime is observed in the wild type or the T534A mutant, suggesting that this is not the only residue modulating substrate selectivity. This is also in agreement with the docking figures of phenylalanine in the CYP79A1 where only a weak contact is observed with the N355, Thirdly, when N-hydroxytyrosine is docked in the active site, the incorporated hydroxyl group interacts with the negatively charged D347 by forming a hydrogen bond thereby fixing its position and preventing rotation around the C-N bond of the N-hydroxytyrosine formed from tyrosine (Fig 2B and 2D). The docking position of N,N-dihydroxytyrosine formed by the second monooxygenation reaction step was stabilized by the same interactions as applied for N-hydroxytyrosine. The N,N-dihydroxytyrosine is highly unstable and rapidly undergoes dehydration via abstraction of the hydrogen atom of the hydroxyl group introduced in the second monooxyganation reaction resulting in the formation of an alpha-nitroso carboxylic acid and 100% retention of the oxygen atom introduced in the second monooxygenation reaction. Finally, decarboxylation via a cyclic transition state leads to the formation of the (E)-p-hydroxyphenylacetaldoxime (Fig 4). The CYP79A1 model obtained thus provides a structural explanation to the biochemical studies demonstrating that the hydroxy group introduced in the second monooxygenation step was 100% retained in the p-hydroxyphenylacetaldoxime formed (35). The mutant D347A displayed only 9.6% of the wild type activity, thus less than the T534A mutant. This residue is not essential for the first hydroxylation step, therefore its mutation may affect the efficiency of the reaction or the stereospecificity of the enzyme but may not be deleterious as the other mutations in conserved positions. Activity assays with N-methyl-tyrosine did not show detectable levels of oximes (Table 4). This analogue has a specific methyl group in the hydroxylation position of tyrosine but maintain the carboxyl and para-hydroxyl group essential for binding. For this reason it is likely that D347 or a similar amino acid along the I´-helix specifically interacts with tyrosine amine group. Consequently a methyl group in this position possibly hinders the specific interaction, thus allowing C-N bond rotation and potentially preventing the hydroxylation cycle. It is worth noticing, that only tyrosine hydroxylating enzymes CYP79A1, CYP79E1 and CYP79E2 show conservation of the aspartate residue in this position, which together with together with the activity of the D347A highlights the importance of this position in the hydroxylation mechanism.

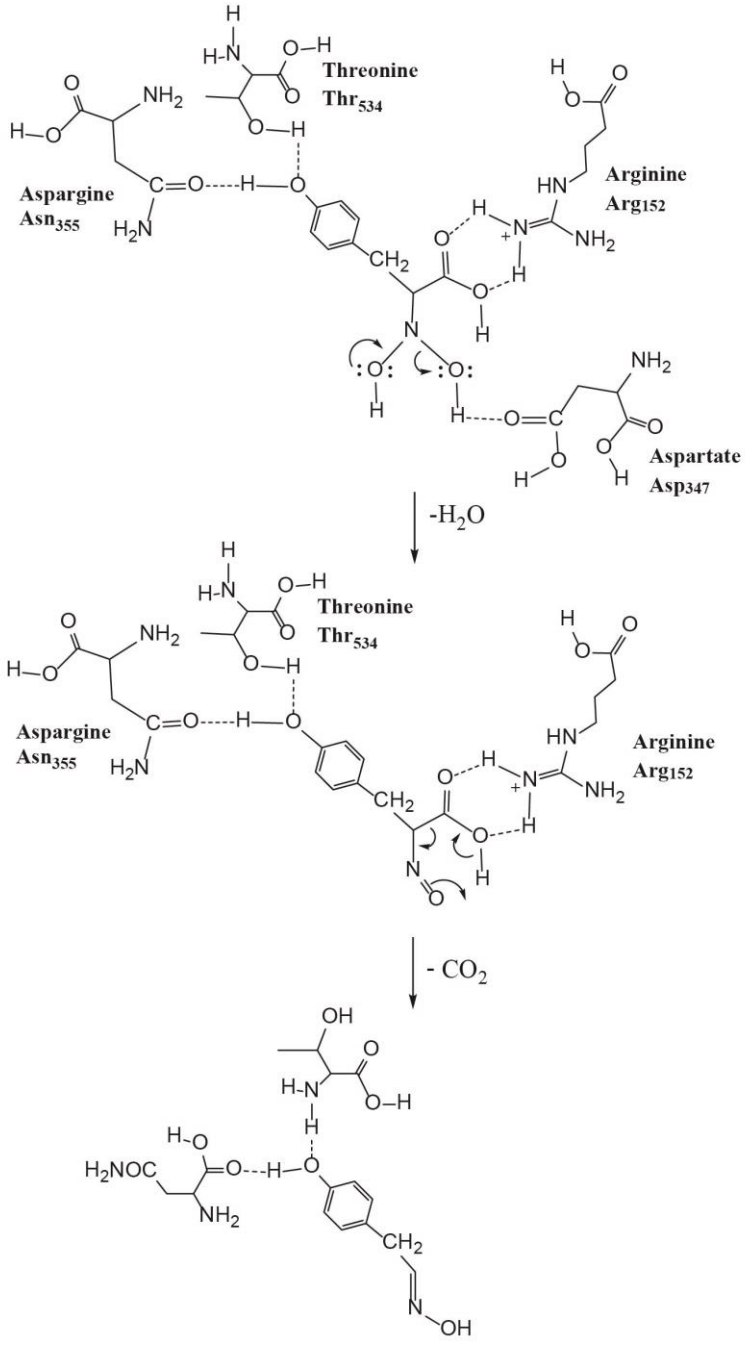


FIG 4. Proposed mechanism for the formation of *(E)-p*-hydroxyphenylacetaldoxime catalyzed by the multifunctional CYP79A1.

In addition to the mutations identified in the docking studies, other mutations in SRS-1, such as E145K, had previously been reported to increase the in planta metabolic flux towards dhurrin formation in sorghum (8). In our in vivo experiments, the E145K mutation resulted in reduced activity. No contacts between E145 and the tyrosine substrate were identified using the current model. The positive effect of this mutation observed

in the plant lines could be explained in several ways. Some mutations make P450s sensitive to an imbalance of reductase equivalents (47). The replacement of a negatively charged residue by a positive could influence important electrostatic interactions on the surface of the enzyme influencing e.g. electron transfer (48). Alternatively, the effect observed in planta could be caused by other not recognized mutations. N355 proved to have a particularly interesting position. The model suggested a direct interaction with the para-hydroxy group of tyrosine, and thereby most likely plays an important role in substrate specificity. However, the closeness to the heme group and the other residues thought to be involved in catalysis could also indicate involvement in the catalytic cycle. The effect of mutations in the SRS-1 (E145K) and BC-loop (R152A) supports the previously suggested important role of this region in determining specificity (16). In this study we also reproduced the mutation P414L located very close to the highly conserved EXXR motif of the P450 family. The lack of activity in this mutant explains the absence of dhurrin in Sorghum plants homozygous for this mutation (8). Multiple alignments also show that this residue is highly conserved in the CYP79 family (Fig. S5).

Given the characteristics of the amino acids participating in the P450 catalyzed hydroxylation cycle, we identified a possible proton delivery and acceptor path within a reasonable distance from the heme group. Our model positions the residues N355 and D354 located in the I-helix at a distance of 3.5Å from the iron plane. N355 may donate a proton to the dioxygen whereas D354 can act as acid-base amino acid replacing the donated proton by the former. Activity assays carried out on intact E. coli cells expressing N355A and D354A confirmed this by showing complete loss-of-function. The alanine substitutions may disrupt the proton donation cycle as these amino acids cannot donate or receive protons. Although these two residues are particularly conserved in all CYP79 enzymes, to our knowledge, there is no prvious mutational data of these residues for this P450 family. However, previous studies in the P450cam demonstrate the role of an aspartate along the I´-helix in the oxygen activation cycle (37). Here for the first time, we provide experimentally validated data for this mechanism for the CYP79 family. Further mutational studies in this region could strengthen this hypothesis. It is worth noting that N355A mutant exhibited lower fluorescence compared to the rest of the mutants. This indicates a possible misfolding of the enzyme giving other clues to the potential mechanism behind the lack of activity.

Conclusions. The new homology modeling strategy here presented provides new insights into the key determinants of CYP79A1 activity and shows that P450s modularity is conserved and can be exploited for guiding the identification of functional regions in cytochrome P450s. This may in turn facilitate a streamlined strategy for engineering enzymes in the absence of structural information. Finally, our findings regarding CYP79A1 shed light on its functionality that can be useful for selection of new crop lines with reduced or depleted production of the toxic cyanogenic glucosides.

Acknowledgements

To Dr. Tomas Laursen for kindly providing purified PORb2 kindly and Helle Munk Petersen for assistance in protein purification. Financial support from the VILLUM Foundation to the research center "Plant Plasticity" and from the UCPH Excellence Program for Interdisciplinary Research to Center of Synthetic Biology "bioSYNergy" is gratefully acknowledged.

Supplementary figures

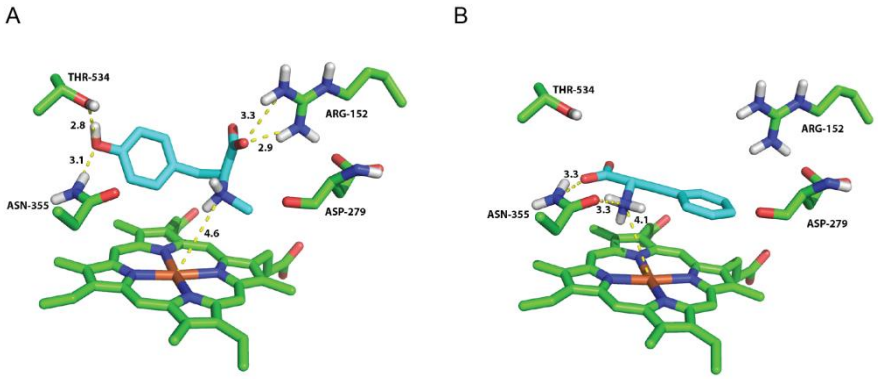


FIG S1 Docking of N-methyl-tyrosine and phenylalanine in the SbCYP79A1 model. N-methyl-tyrosine is shown in the same binding mode as tyrosine in the active, with R152 interacting with the carboxylic group of tyrosine at a distance of 2.9Å and the para-hydroxyl group stabilized by T534 at 2.8Å, No contacts are observed with the D347 (A). Phenylalanine docked in the active site parallel to the heme group showing only one contact at a distance of 3.3Å distance from the amino acid residue N355 (B).

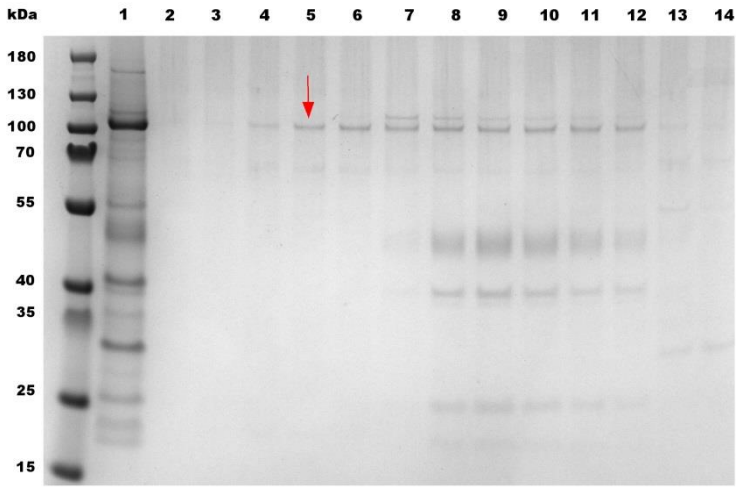


FIG S2 SDS-PAGE Analysis of the *dsbA*-sbCYP79A1-*GFP* purification showing solubilized membrane extract from *E. coli*, washed fraction and elution fractions 1 to 12 from gel filtration purification. The red arrow points the band corresponding to the purified protein.

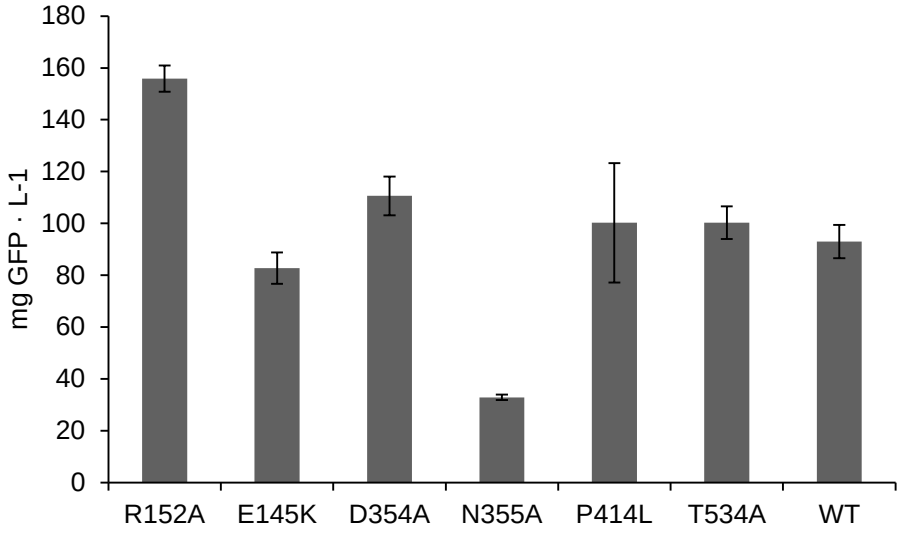


FIG S3 Expression levels calculated in mg/L of GFP fusions of *Sb*CYP79A1 mutants and control (non-mutated) in *E .coli*. The data corresponds to average and standard deviation of three biological replicates.

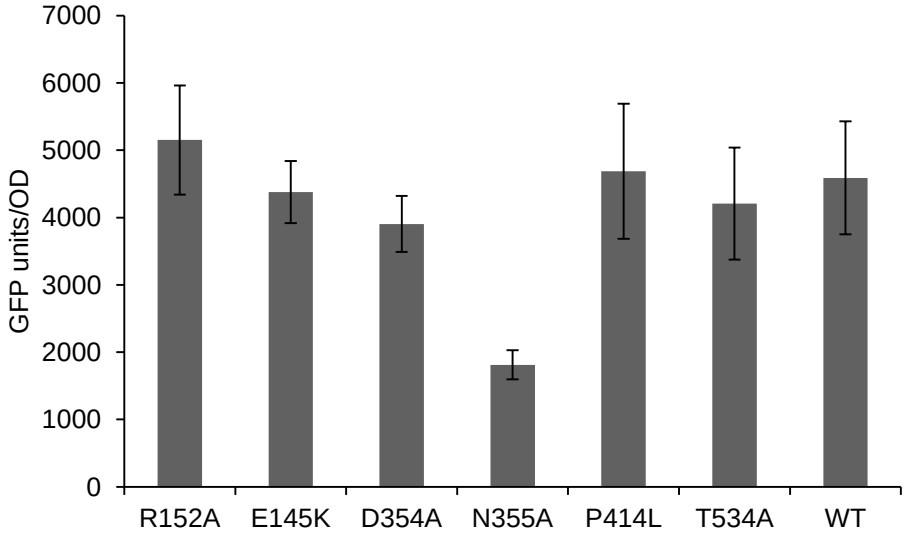


FIG S4 Corrected GFP activity by optical density of expressed *Sb*CYP79A1mutants and wild type in *E. coli*. The data corresponds to average and standard deviation of three biological replicates.

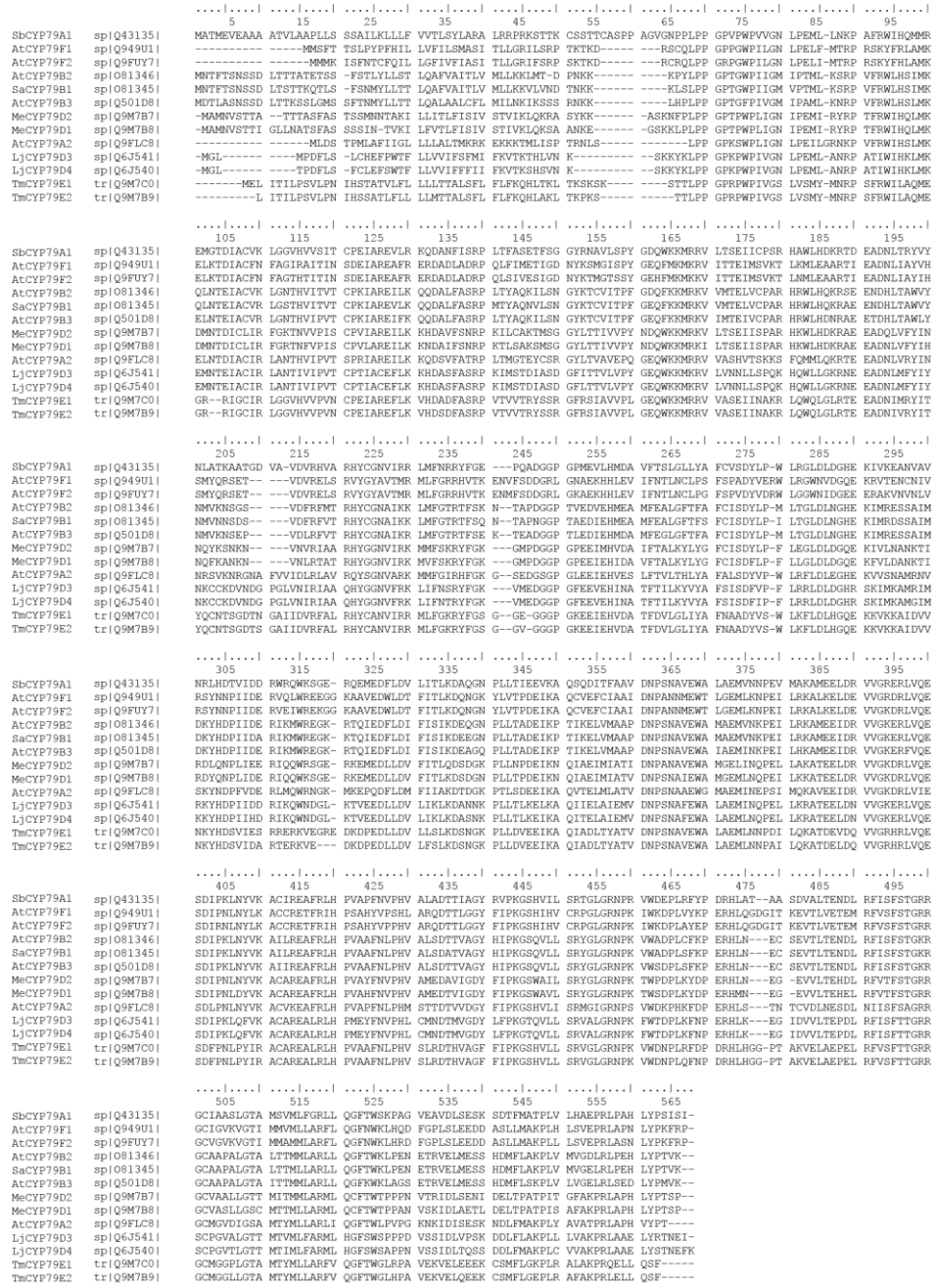


FIG S5 Multiple alignment of the characterized CYP79 family enzymes with ClustalW. Sb: *Sorghum bicolor*, Sa: *Sinapis alba*, At: *Arabidopsis thaliana*, Tm: *Triglochin maritima*, Me: *Manihot esculenta*, Lj: *Lotus japonicus*

References

1. Neilson EH, Goodger JQ, Woodrow IE, Moller BL. 2013. Plant chemical defense: at what cost? Trends Plant Sci 18:250-258.

2. Podust L, Sherman D. 2012. Diversity of P450 enzymes in the biosynthesis of natural products. Natural product reports 29:1251-1266.

3. Morant M, Bak S, Moller BL, Werck-Reichhart D. 2003. Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. Curr Opin Biotechnol 14:151-162.

4. Bak S, Beisson F, Bishop G, Hamberger B, Hofer R, Paquette S, Werck-Reichhart D. 2011. Cytochromes p450. Arabidopsis Book 9:e0144.

5. Graham SE, Peterson JA. 1999. How similar are P450s and what can their differences teach us? Arch Biochem Biophys 369:24-29.

6. Rupasinghe S, Schuler MA, Kagawa N, Yuan H, Lei L, Zhao B, Kelly SL, Waterman MR, Lamb DC. 2006. 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 580:6338-6342.

7. Moller BL. 2010. Functional diversifications of cyanogenic glucosides. Curr Opin Plant Biol 13:338-347.

8. Blomstedt CK, Gleadow RM, O'Donnell N, Naur P, Jensen K, Laursen T, Olsen CE, Stuart P, Hamill JD, Moller BL, Neale AD. 2012. A combined biochemical screen and TILLING approach identifies mutations in Sorghum bicolor L. Moench resulting in acyanogenic forage production. Plant Biotechnol J 10:54-66.

9. Clausen M, Kannangara RM, Olsen CE, Blomstedt CK, Gleadow RM, Jorgensen K, Bak S, Motawie MS, Moller BL. 2015. The bifurcation of the cyanogenic glucoside and glucosinolate biosynthetic pathways. Plant J 84:558-573.

10. Irmisch S, McCormick AC, Boeckler GA, Schmidt A, Reichelt M, Schneider B, Block K, Schnitzler JP, Gershenzon J, Unsicker SB, Kollner TG. 2013. Two herbivore-induced cytochrome P450 enzymes CYP79D6 and CYP79D7 catalyze the formation of volatile aldoximes involved in poplar defense. Plant Cell 25:4737-4754.

11. Naur P, Hansen CH, Bak S, Hansen BG, Jensen NB, Nielsen HL, Halkier BA. 2003. CYP79B1 from Sinapis alba converts tryptophan to indole-3-acetaldoxime. Arch Biochem Biophys 409:235-241.

12. Hull AK, Vij R, Celenza JL. 2000. Arabidopsis cytochrome P450s that catalyze the first step of tryptophan-dependent indole-3-acetic acid biosynthesis. Proc Natl Acad Sci U S A 97:2379-2384.

13. Andersen MD, Busk PK, Svendsen I, Moller BL. 2000. Cytochromes P-450 from cassava (Manihot esculenta Crantz) catalyzing the first steps in the biosynthesis of the cyanogenic glucosides linamarin and lotaustralin. Cloning, functional expression in Pichia pastoris, and substrate specificity of the isolated recombinant enzymes. J Biol Chem 275:1966-1975.

14. Nielsen JS, Moller BL. 2000. Cloning and expression of cytochrome P450 enzymes catalyzing the conversion of tyrosine to p-hydroxyphenylacetaldoxime in the biosynthesis of cyanogenic glucosides in Triglochin maritima. Plant Physiol 122:1311-1321.

15. Sibbesen O, Koch B, Halkier BA, Moller BL. 1995. Cytochrome P-450TYR is a multifunctional heme-thiolate enzyme catalyzing the conversion of L-tyrosine to p-hydroxyphenylacetaldehyde oxime in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench. J Biol Chem 270:3506-3511.

16. Kahn RA, Fahrendorf T, Halkier BA, Moller BL. 1999. Substrate specificity of the cytochrome P450 enzymes CYP79A1 and CYP71E1 involved in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench. Arch Biochem Biophys 363:9-18.

17. Jensen K, Osmani SA, Hamann T, Naur P, Moller BL. 2011. Homology modeling of the three membrane proteins of the dhurrin metabolon: catalytic sites, membrane surface association and protein-protein interactions. Phytochemistry 72:2113-2123.

18. Baudry J, Rupasinghe S, Schuler MA. 2006. Class-dependent sequence alignment strategy improves the structural and functional modeling of P450s. Protein Eng Des Sel 19:345-353.

19. Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J. 1995. Structure and function of cytochromes P450: a comparative analysis of three crystal structures. Structure 3:41-62.

20. Rubtsov P, Nizhnik A, Dedov I, Kalinchenko N, Petrov V, Orekhova A, Spirin P, Prassolov V, Tiulpakov A. 2015. Partial deficiency of 17alpha-hydroxylase/17,20-lyase caused by a novel missense mutation in the canonical cytochrome heme-interacting motif. Eur J Endocrinol 172:K19-25.

21. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, Honigschmid P, Schafferhans A, Roos M, Bernhofer M, Richter L, Ashkenazy H, Punta M, Schlessinger A, Bromberg Y, Schneider R, Vriend G, Sander C, Ben-Tal N, Rost B. 2014. PredictProtein--an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 42:W337-343.

22. Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779-815.

23. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25:1605-1612.

24. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT. 2006. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J Med Chem 49:6177-6196.

25. Hamberger B, Bak S. 2013. Plant P450s as versatile drivers for evolution of species-specific chemical diversity. Philos Trans R Soc Lond B Biol Sci 368:20120426.

26. Cavaleiro AM, Kim SH, Seppala S, Nielsen MT, Norholm MH. 2015. Accurate DNA Assembly and Genome Engineering with Optimized Uracil Excision Cloning. ACS Synth Biol 4:1042-1046.

27. Norholm MH. 2010. A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. BMC Biotechnol 10:21.

28. Drew DE, von Heijne G, Nordlund P, de Gier JW. 2001. Green fluorescent protein as an indicator to monitor membrane protein overexpression in Escherichia coli. FEBS Lett 507:220-224.

29. Drew D, Lerch M, Kunji E, Slotboom DJ, de Gier JW. 2006. Optimization of membrane protein overexpression and purification using GFP fusions. Nat Methods 3:303-313.

30. Mirzadeh K, Martinez V, Toddo S, Guntur S, Herrgard MJ, Elofsson A, Norholm MH, Daley DO. 2015. Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector-Coding Sequence Junction. ACS Synth Biol doi:10.1021/acssynbio.5b00033.

31. Laursen T, Jensen K, Moller BL. 2011. Conformational changes of the NADPH-dependent cytochrome P450 reductase in the course of electron transfer to cytochromes P450. Biochim Biophys Acta 1814:132-138.

32. Moller BL, Conn EE. 1979. The biosynthesis of cyanogenic glucosides in higher plants. N-Hydroxytyrosine as an intermediate in the biosynthesis of dhurrin by Sorghum bicolor (Linn) Moench. J Biol Chem 254:8575-8583.

33. Halkier BA, Olsen CE, Moller BL. 1989. The biosynthesis of cyanogenic glucosides in higher plants. The (E)- and (Z)-isomers of p-hydroxyphenylacetaldehyde oxime as intermediates in the biosynthesis of dhurrin in Sorghum bicolor (L.) Moench. J Biol Chem 264:19487-19494.

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

35. Halkier BA, Lykkesfeldt J, Moller BL. 1991. 2-nitro-3-(p-hydroxyphenyl)propionate and aci-1-nitro-2-(p-hydroxyphenyl)ethane, two intermediates in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench. Proc Natl Acad Sci U S A 88:487-491.

36. Meunier B, de Visser SP, Shaik S. 2004. Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes. Chem Rev 104:3947-3980.

37. Vidakovic M, Sligar SG, Li H, Poulos TL. 1998. Understanding the role of the essential Asp251 in cytochrome p450cam using site-directed mutagenesis, crystallography, and kinetic solvent isotope effect. Biochemistry 37:9211-9219.

38. Zheng M, Luo X, Shen Q, Wang Y, Du Y, Zhu W, Jiang H. 2009. Site of metabolism prediction for six biotransformations mediated by cytochromes P450. Bioinformatics 25:1251-1258.

39. Seger ST, Rydberg P, Olsen L. 2015. Mechanism of the N-hydroxylation of primary and secondary amines by cytochrome P450. Chem Res Toxicol 28:597-603.

40. Mikkelsen MD, Hansen CH, Wittstock U, Halkier BA. 2000. Cytochrome P450 CYP79B2 from Arabidopsis catalyzes the conversion of tryptophan to indole-3-acetaldoxime, a precursor of indole glucosinolates and indole-3-acetic acid. J Biol Chem 275:33712-33717.

41. Schierle CF, Berkmen M, Huber D, Kumamoto C, Boyd D, Beckwith J. 2003. The DsbA signal sequence directs efficient, cotranslational export of passenger proteins to the Escherichia coli periplasm via the signal recognition particle pathway. J Bacteriol 185:5706-5713.

42. Cavasotto CN, Phatak SS. 2009. Homology modeling in drug discovery: current trends and applications. Drug Discov Today 14:676-683.

43. Halkier BA, Moller BL. 1990. The biosynthesis of cyanogenic glucosides in higher plants. Identification of three hydroxylation steps in the biosynthesis of dhurrin in Sorghum bicolor (L.) Moench and the involvement of 1-ACI-nitro-2-(p-hydroxyphenyl)ethane as an intermediate. J Biol Chem 265:21114-21121.

44. Roberts AG, Cheesman MJ, Primak A, Bowman MK, Atkins WM, Rettie AE. 2010. Intramolecular heme ligation of the cytochrome P450 2C9 R108H mutant demonstrates pronounced conformational flexibility of the B-C loop region: implications for substrate binding. Biochemistry 49:8700-8708.

45. Chen JS, Berenbaum MR, Schuler MA. 2002. Amino acids in SRS1 and SRS6 are critical for furanocoumarin metabolism by CYP6B1v1, a cytochrome P450 monooxygenase. Insect Mol Biol 11:175-186.

46. Wittstock U, Halkier BA. 2000. Cytochrome P450 CYP79A2 from Arabidopsis thaliana L. Catalyzes the conversion of L-phenylalanine to phenylacetaldoxime in the biosynthesis of benzylglucosinolate. J Biol Chem 275:14659-14666.

47. Kaspera R, Naraharisetti SB, Evangelista EA, Marciante KD, Psaty BM, Totah RA. 2011. Drug metabolism by CYP2C8.3 is determined by substrate dependent interactions with cytochrome P450 reductase and cytochrome b5. Biochem Pharmacol 82:681-691.

48. Zhao C, Gao Q, Roberts AG, Shaffer SA, Doneanu CE, Xue S, Goodlett DR, Nelson SD, Atkins WM. 2012. Cross-linking mass spectrometry and mutagenesis confirm the functional importance of surface interactions between CYP3A4 and holo/apo cytochrome b(5). Biochemistry 51:9488-9500

CHAPTER 4

Back-up DNA technologies for optimizing cytochrome P450 expression

Randomization of the vector-coding sequence junction improves expression of the plant cytochrome P450 CYP79A1

BACKGROUND

N-terminal modifications of P450s have proven to be successful in improving P450 expression to different degrees in E. coli. In chapter 2 we described several N-terminal tags with positive impact on P450 expression possibly through different mechanisms such as improved targeting to the SRP-dependent pathway, the endogenous insertion machinery or increased mRNA stability. In some cases these modifications may not be enough to achieve satisfactory amounts of P450s for a certain applications. The so-called 28 amino acid tag likely works by minimizing the mRNA folding energy, thus facilitating translationional initiation (Kudla, Murray et al. 2009). The fusion of this short amino acid sequence to the N-termini of CYP79A1 and CYP71E1 resulted in improved functional expression of both enzymes. Although codon usage in the N-termini could also have an effect, the tag also had a positive impact on a library of codon-optimized P450s. Furthermore, previous evidences suggest that AT-rich sequences and alanine substitutions in the N-terminus of P450s may result in higher P450 expression (Zelasko, Palaria et al. 2013). Daley and co-workers observed large variations of gene expression levels caused by different restriction enzyme scars in a region immediately upstream from the start codon (Mirzadeh, Martinez et al. 2015). (Mirzadeh, Martinez et al. 2015). This led to a new and simple way to optimize gene expression, by tuning the translation-initiation region, and more specifically the vector-coding sequence junction between the Shine-Dalgarno sequence and the two first codons downstream from ATG. Combining a complete randomization of the six nucleotides in the upstream sequence of the start codon with synonymous mutations of these two codons, it was possible to improve expression of E. coli membrane proteins up to 1000 fold (Mirzadeh, Martinez et al. 2015). Inspired by these studies, we here apply the vector:coding sequence optimization strategy to improve the expression of membrane-associated plant P450s, expanding on the toolbox of techniques to facilitate the development of P450 biocatalysts.

MATERIALS AND METHODS

Randomization of the vector-encoding region of the pET28a(+)-CYP79A1 derived construct

One fragment uracil excision cloning was performed as previously described (Cavaleiro, Kim et al. 2015)chapters using a pET28a(+) derived vector containing the open reading frame of the CYP79A1 followed by a TEV cleavage site, a GFP fusion and a hexa histidine tail. Desired mutations were introduced in the reverse primer; 1) Six nucleotides preceding the ATG were completely randomized and 2) synonymous alanine mutations in the two first codons of the gene (**Fig. 21**). PCR products were treated with DpnI for 2h at 37°C and gel purified. Purified PCR products were then incubated with 1 µL USER enzyme (New England BioLabs, Ipswich, USA) for 30 min at 37°C and at 20°C for 15 min. The whole uracil-excision reaction was transformed into BL21(DE3) chemically competent cells according to the manufacturer's protocol. After recovery 1/5 of cells were plated on Luria Bertoni (LB) agar supplemented with 50 µg/mL kanamycin. Remaining cells were transferred to pre-expression cultures. Colonies grown on plates were screened by DNA sequencing in order to determine clone diversity (Eurofins Genomics, Ebersberg, Germany).

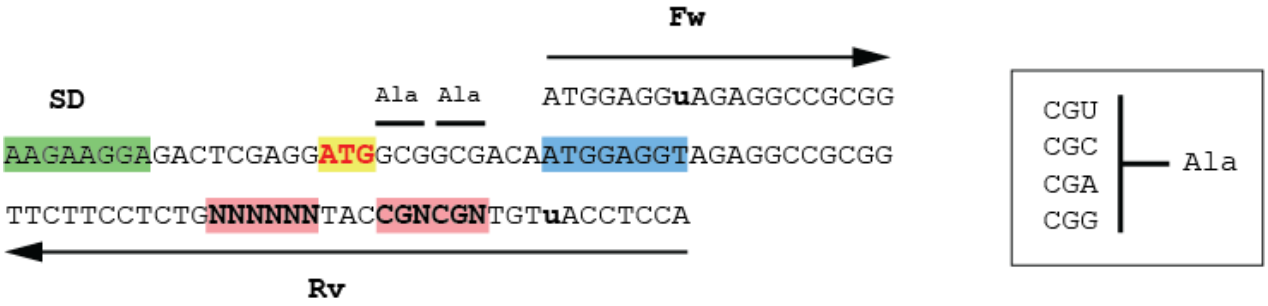


Fig. 21 Randomization of the vector-gene encoding junction of the CYP79A1 symbolized by the letter **N**. In green the Shine Delgarno sequence (SD), in red boxes the mutations introduced in the reverse primer (Rv) flanking the ATG initiation codon, and in the blue box the uracil-excision splitting point to generate complementary overhangs in the primers (Fw and Rv). On the left box synonymous codons for alanine are shown.

Expression of the randomized CYP79A1 library

After transformation BL21(DE3) cells were inoculated in LB media supplemented with 100 µg/mL kanamycin in 5ml liquid cultures and grown over-night. Expression cultures were initiated at a final OD of 0.05 in 5ml of fresh LB medium. Absorbance measurements were carried out as described in the previous chapters at 600nm in a SynergyMx, SMATLD plate reader (BioTek, Winooski, USA). Growth was performed for approximately 2 h at 30°C, 250 rpm in an Innova®44R incubator shaker system (5 cm orbital

shaking) (New Brunswick Scientific, Eppendorf, USA). Protein expression was induced at 0.5 OD with IPTG (isopropyl ß-D-1-thiogalactopyranoside, dioxane free, Thermo Scientific, Waltham, USA) at a final concentration of 0.5mM. Expression was allowed for 5h at 25°C and 250rpm.

Fluorescence measurements

Single cell expression was determined by flow cytometry with a BD LSRFortessa™ flow cytometer. Cells were diluted 1/200 in PBS, incubated for 15 min at room temperature and green fluorescence was collected in the FITC channel (488 nm emission wavelength). Expression was benchmarked to BL21(DE3) cells containing the empty vector, the 28 amino acid tag, and the native CYP79A1 fusion construct. Flow cytometry data was exported in FCS files and analysed using the *flowViz* and *flowcore* packages with R software as previously described (http://www.R-project.org/).

RESULTS

Following nucleotide randomization five different colonies were selected for DNA sequencing. The analysis revealed that four out of the five clones analyzed displayed differences in the sequence (**Fig. 22**). However, only ten colonies were obtained in 1/5 of the library. The CYP79A1 randomized library was expressed in BL21(DE3) and fluorescence measured after 5h induction by flow cytometry. Results clearly show the positive impact of randomization on expression of the CYP79A1 (**Fig. 22**). While the wild type CYP79A1 is not expressed as compared to empty vector cells, the randomized CYP79A1 library stands above these values. Additionally the 28 amino acid tagged CYP79A1 display higher average expression than the randomized library. Nevertheless the best expressers of the randomized library display as much fluorescence as the average of the 28 tagged CYP79A1 (**Fig. 22**).

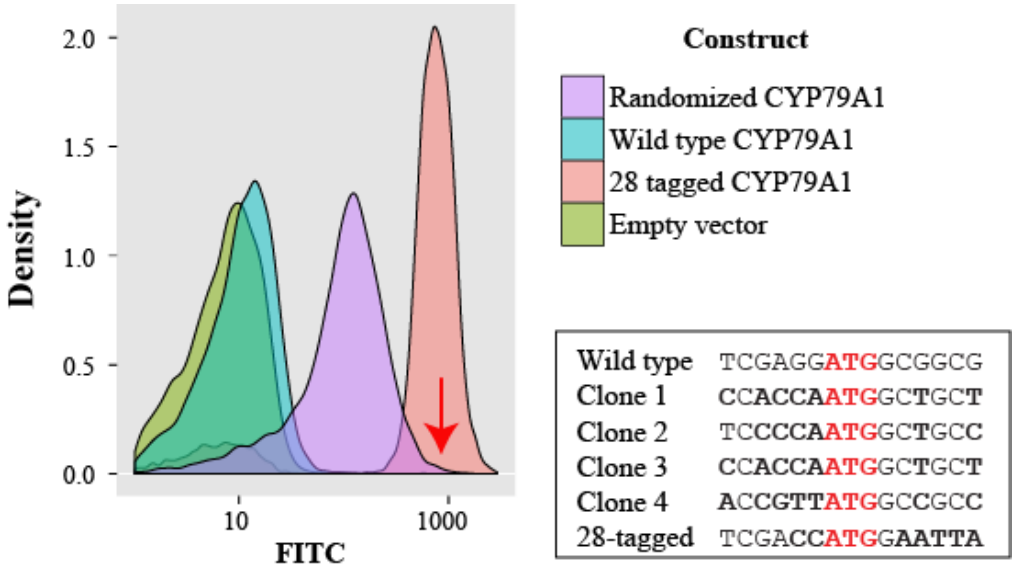


Fig. 22 Expression analysis of the different constructs by flow cytometry. The red arrow points the best expressers of the randomized CYP79A1 library. The bottom right box shows the sequencing results of the randomized library with mutated nucleotides in bold.

DISCUSSION

The success of production of natural compounds in microbial hosts relies often on the last steps of biosynthetic pathways carried out by P450s. We have previously expanded the toolbox of N-terminal tags to allow for functional expression P450 enzymes in *E. coli* using a GFP-based platform. Here we use a previously unexplored approach for P450 expression. With a very simple PCR-based strategy, the six nucleotides upstream from the ATG start codon together with synonymous mutations of the CYP79A1 have been randomized. Thus, keeping the wild type peptide sequence without the need for *de novo* codon-optimization of the entire CYP79A1 coding sequence. We have specifically verified the effect of this randomization strategy using the GFP platform described in the previous chapters and benchmarked it to the 28-tagged CYP79A1.

The results of the sequencing analysis have shown that a one-fragment uracil-excision cloning generates DNA diversity in the targeted region. Unfortunately the size of the library is still limited to a dozen of clones judging by the number of colonies obtained after plating 1/5 of the library. However, we have to bear in mind that only a small fraction of the uracil-excision cloning reaction is used for further transformation. This could be one of the reasons for which the final yield is rather low. Pooling individual reactions in bigger transformation batches or scale-up overnight transformation cultures could result in higher yields.

Finally we have expressed the randomized library of the CYP79A1 along the 28-tagged monitoring fluorescence by flow cytometry. The randomization of the vector-gene coding region resulted in an increased of expression compared to the wild type sequence. Although the randomized library do not outperform the 28-tagged CYP79A1 it shows the potential of the technique to improve P450 expression. The population analysis by flow cytometry also reveals a small fraction of high expressers in the range of the 28-tagged CYP78A1 that could be sorted by FACS and sequenced as previously described. In this study we have carried out only a proof-of-concept expression experiment to explore alternative technologies that deal with underlying biophysical parameters of nucleic acid thermodynamics. The factors influencing expression on DNA level are still unclear though, previous observations by Daley and co-workers point out towards RNA folding energies (Mirzadeh, Martinez et al. 2015). Unfortunately this factor cannot always explain the effects of sequence composition in expression. Nevertheless vector-gene coding junction randomization can provide a back-up strategy when N-terminal tags may fail or be insufficient.

Concluding remarks and future perspectives

The ultimate goal of cell factory development is the production of high value chemicals from renewable feedstocks to reduce our dependency on oil derivatives. But most importantly, many natural chemicals are simply too difficult to produce by the traditional organic synthesis technology. Plant cytochrome P450 enzymes are at the core of natural chemicals since they are involved in almost all biosynthetic pathways leading to high value compounds. Some of these compounds have been mentioned in the introductory part of this thesis and are of high importance because they can be used as medicines. The antimalarial artemisinin or antineoplastic taxol represent the pragmatic examples of P450-dependent chemicals. However, after decades of intensive research in P450s, expression of these enzymes in microbes still represents the main bottleneck to their potential use as biocatalysts.

To solve some of the technical challenges in P450 expression and engineering several strategies are available. On one hand the membrane protein and crystallography field have greatly contributed to the development of bacterial strains able to cope with toxicity of hydrophobic proteins. Generally these strains provide certain control to attenuate saturation of the protein translocation machinery of the host. Engineered promoters, or plasmid-based systems capable of reducing the amount of T7 RNA polymerase, have been successful in facilitating heterologous expression of membrane protein. On the other hand N-terminal modification of P450s represent a wide spread and empirical approach to improve expression. By either facilitating the recognition by the host translocation machinery, targeting to the proper membrane compartment or mRNA stability, these modifications serve as starting point for P450 cell factory engineering. Unfortunately the myriad of strategies available, makes it difficult to choose wisely the tools for successful engineering a functional P450 variant.

In the chapter 1 we have developed a GFP-based optimization platform in *E. coli* that, at least, allows for systematic workflow to analyze P450 expression, putting aside techniques traditionally used to characterize P450s. This platform relies on a GFP reporter that fluoresce on the cytoplasmic side of the cell if the fusion protein is properly folded. Using this approach we have been able to find conditions, strains, and strategies suitable to improve P450 expression. The study also discusses the high throughput limitations of P450 transmembrane region truncations to improve expression of new P450s.

In the chapter 2 we specifically exploit this platform to expand the number of N-terminal modifications available to the date analyzing their effect in two strains, and provide evidence of their usefulness in a metabolic engineering scenario. Finally, the effect of a short 28 amino acids sequence for improving expression of a large library of medicinal plant P450 enzymes is tested. The results from this large screening demonstrate that this tag can improve expression of nearly half of the P450s of the library by more than 2 fold.

In the chapter 3 we use the knowledge generated during the previous chapters to elucidate the structure of cytochromes P450 using homology-based modeling. With these two techniques we have been able to identify important regions for substrate binding and catalysis of CYP79A1, which has not been yet crystallized. The structural characterization of CYP79A1 not only revealed a possible reaction mechanism of the N-hydroxylation of tyrosine by this enzyme but also shed light into the conservation of residues in the CYP79 family involved in the biosynthesis of the plant defense compounds called cyanogenic glucosides.

Finally, based on previous data with the short 28 amino acid tag, and studies carried out by Daley and co-workers we decided to adapt a vector optimization apprach for P450 expression. In the chapter 4 we report the randomization of the vector-gene coding junction using uracil-excision cloning. Using this simple DNA editing technique is possible to improve expression of the cytochrome P450 CYP79A1. This was the first time that this technique has been applied successfully to plant P450s.

All these technologies taken together can serve as starting point to optimize new plant P450s involved in the biosynthesis of high valuable compounds. One example of this is a highly expressed P450s CYP726A1 tested in the chapter 2. This family of P450s found in *Euphorbia peplus* is thought to be involved in the metabolic pathway leading to the formation of ingenol. This compound serves as treatment against actinic keratosis, a skin disease. Thus, the technologies developed in this thesis could greatly contribute to the reconstruction of plant metabolic pathways in *E. coli*, and ultimately help to develop robust cell factories.

References

Ajikumar, P. K., W. H. Xiao, K. E. Tyo, Y. Wang, F. Simeon, E. Leonard, O. Mucha, T. H. Phon, B. Pfeifer and G. Stephanopoulos (2010). "Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli." Science **330**(6000): 70-74.

Almen, M. S., K. J. Nordstrom, R. Fredriksson and H. B. Schioth (2009). "Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin." BMC Biol **7**: 50.

Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner and H. Mori (2006). "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection." Mol Syst Biol **2**: 2006 0008.

Bak, S., R. A. Kahn, H. L. Nielsen, B. L. Moller and B. A. Halkier (1998). "Cloning of three A-type cytochromes P450, CYP71E1, CYP98, and CYP99 from Sorghum bicolor (L.) Moench by a PCR approach and identification by expression in Escherichia coli of CYP71E1 as a multifunctional cytochrome P450 in the biosynthesis of the cyanogenic glucoside dhurrin." Plant Mol Biol **36**(3): 393-405.

Bak, S., C. E. Olsen, B. A. Halkier and B. L. Moller (2000). "Transgenic tobacco and Arabidopsis plants expressing the two multifunctional sorghum cytochrome P450 enzymes, CYP79A1 and CYP71E1, are cyanogenic and accumulate metabolites derived from intermediates in Dhurrin biosynthesis." Plant Physiol **123**(4): 1437-1448.

Baker, D. and A. Sali (2001). "Protein structure prediction and structural genomics." Science **294**(5540): 93-96.

Barnes, H. J., M. P. Arlotto and M. R. Waterman (1991). "Expression and enzymatic activity of recombinant cytochrome P450 17 alpha-hydroxylase in Escherichia coli." Proc Natl Acad Sci U S A **88**(13): 5597-5601.

Baumler, D. J., R. G. Peplinski, J. L. Reed, J. D. Glasner and N. T. Perna (2011). "The evolution of metabolic networks of E. coli." BMC Syst Biol **5**: 182.

Belin, D. (2010). "In vivo analysis of protein translocation to the Escherichia coli periplasm." Methods Mol Biol **619**: 103-116.

Bogdanov, M., W. Zhang, J. Xie and W. Dowhan (2005). "Transmembrane protein topology mapping by the substituted cysteine accessibility method (SCAM(TM)): application to lipid-specific membrane protein topogenesis." Methods **36**(2): 148-171.

Budriang, C., P. Rongnoparut and J. Yuvaniyama (2011). "An expression of an insect membrane-bound cytochrome P450 CYP6AA3 in the Escherichia coli in relation to insecticide resistance in a malarial vector." Pak J Biol Sci **14**(8): 466-475.

Burley, S. K., S. C. Almo, J. B. Bonanno, M. Capel, M. R. Chance, T. Gaasterland, D. Lin, A. Sali, F. W. Studier and S. Swaminathan (1999). "Structural genomics: beyond the human genome project." Nat Genet **23**(2): 151-157.

Carpenter, E. P., K. Beis, A. D. Cameron and S. Iwata (2008). "Overcoming the challenges of membrane protein crystallography." Curr Opin Struct Biol **18**(5): 581-586.

Casini, A., M. Storch, G. S. Baldwin and T. Ellis (2015). "Bricks and blueprints: methods and standards for DNA assembly." Nat Rev Mol Cell Biol **16**(9): 568-576.

Cavaleiro, A. M., S. H. Kim, S. Seppala, M. T. Nielsen and M. H. Norholm (2015). "Accurate DNA Assembly and Genome Engineering with Optimized Uracil Excision Cloning." ACS Synth Biol **4**(9): 1042-1046.

Chang, M. C., R. A. Eachus, W. Trieu, D. K. Ro and J. D. Keasling (2007). "Engineering Escherichia coli for production of functionalized terpenoids using plant P450s." Nat Chem Biol **3**(5): 274-277.

Chang, M. C. and J. D. Keasling (2006). "Production of isoprenoid pharmaceuticals by engineered microbes." Nat Chem Biol **2**(12): 674-681.

Chang, Z., X. Wang, R. Wei, Z. Liu, H. Shan, G. Fan and H. Hu (2010). "Functional Expression and Purification of CYP93C20a Plant Membrane-Associated Cytochrome P450 from Medicago truncatula." Protein Expr Purif.

Chauhan, R., R. Jones, P. Wilkinson, Y. Pauchet and R. H. Ffrench-Constant (2013). "Cytochrome P450-encoding genes from the Heliconius genome as candidates for cyanogenesis." Insect Mol Biol **22**(5): 532-540.

Daley, D. O., M. Rapp, E. Granseth, K. Melen, D. Drew and G. von Heijne (2005). "Global topology analysis of the Escherichia coli inner membrane proteome." Science **308**(5726): 1321-1323.

Danielson, P. B. (2002). "The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans." Curr Drug Metab **3**(6): 561-597.

de Marco, A. (2009). "Strategies for successful recombinant expression of disulfide bond-dependent proteins in Escherichia coli." Microb Cell Fact **8**: 26.

De Marothy, M. T. and A. Elofsson (2015). "Marginally hydrophobic transmembrane alpha-helices shaping membrane protein folding." Protein Sci **24**(7): 1057-1074.

Demain, A. L. (2006). "From natural products discovery to commercialization: a success story." J Ind Microbiol Biotechnol **33**(7): 486-495.

Denks, K., A. Vogt, I. Sachelaru, N. A. Petriman, R. Kudva and H. G. Koch (2014). "The Sec translocon mediated protein transport in prokaryotes and eukaryotes." Mol Membr Biol **31**(2-3): 58-84.

Drew, D., M. Lerch, E. Kunji, D. J. Slotboom and J. W. de Gier (2006). "Optimization of membrane protein overexpression and purification using GFP fusions." Nat Methods **3**(4): 303-313.

Drew, D., S. Newstead, Y. Sonoda, H. Kim, G. von Heijne and S. Iwata (2008). "GFP-based optimization scheme for the overexpression and purification of eukaryotic membrane proteins in Saccharomyces cerevisiae." Nat Protoc **3**(5): 784-798.

Drew, D. E., G. von Heijne, P. Nordlund and J. W. de Gier (2001). "Green fluorescent protein as an indicator to monitor membrane protein overexpression in Escherichia coli." FEBS Lett **507**(2): 220-224.

Durante-Rodriguez, G., V. de Lorenzo and E. Martinez-Garcia (2014). "The Standard European Vector Architecture (SEVA) plasmid toolkit." Methods Mol Biol **1149**: 469-478.

Egan, S. M. and R. F. Schleif (1993). "A regulatory cascade in the induction of rhaBAD." J Mol Biol **234**(1): 87-98.

Eiteman, M. A. and E. Altman (2006). "Overcoming acetate in Escherichia coli recombinant protein fermentations." Trends Biotechnol **24**(11): 530-536.

Exposito, O., M. Bonfill, E. Moyano, M. Onrubia, M. H. Mirjalili, R. M. Cusido and J. Palazon (2009). "Biotechnological production of taxol and related taxoids: current state and prospects." Anticancer Agents Med Chem **9**(1): 109-121.

Feilmeier, B. J., G. Iseminger, D. Schroeder, H. Webber and G. J. Phillips (2000). "Green fluorescent protein functions as a reporter for protein localization in Escherichia coli." J Bacteriol **182**(14): 4068-4076.

Fernandez, C. and K. Wuthrich (2003). "NMR solution structure determination of membrane proteins reconstituted in detergent micelles." FEBS Lett **555**(1): 144-150.

Frank, K. and M. J. Sippl (2008). "High-performance signal peptide prediction based on sequence alignment techniques." Bioinformatics **24**(19): 2172-2176.

Fujiyoshi, Y. (2011). "Electron crystallography for structural and functional studies of membrane proteins." J Electron Microsc (Tokyo) **60 Suppl 1**: S149-159.

Giacalone, M., A. Gentile, B. Lovitt, N. Berkley, C. Gunderson and M. Surber (2006). "Toxic protein expression in Escherichia coli using a rhamnose-based tightly regulated and tunable promoter system." BioTechniques **40**(3): 355-364.

Giacalone, M. J., A. M. Gentile, B. T. Lovitt, N. L. Berkley, C. W. Gunderson and M. W. Surber (2006). "Toxic protein expression in Escherichia coli using a rhamnose-based tightly regulated and tunable promoter system." Biotechniques **40**(3): 355-364.

Gillam, E. M., R. M. Wunsch, Y. F. Ueng, T. Shimada, P. E. Reilly, T. Kamataki and F. P. Guengerich (1997). "Expression of cytochrome P450 3A7 in Escherichia coli: effects of 5' modification and catalytic characterization of recombinant enzyme expressed in bicistronic format with NADPH-cytochrome P450 reductase." Arch Biochem Biophys **346**(1): 81-90.

Gnanasekaran, T., K. Vavitsas, J. Andersen-Ranberg, A. Z. Nielsen, C. E. Olsen, B. Hamberger and P. E. Jensen (2015). "Heterologous expression of the isopimaric acid pathway in Nicotiana benthamiana and the effect of N-terminal modifications of the involved cytochrome P450 enzyme." J Biol Eng **9**: 24.

Goldman, B. M. and G. Blobel (1981). "In vitro biosynthesis, core glycosylation, and membrane integration of opsin." J Cell Biol **90**(1): 236-242.

Gotzke, H., C. Muheim, A. F. Altelaar, A. J. Heck, G. Maddalo and D. O. Daley (2015). "Identification of putative substrates for the periplasmic chaperone YfgM in Escherichia coli using quantitative proteomics." Mol Cell Proteomics **14**(1): 216-226.

Hahne, F., N. LeMeur, R. R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain and R. Gentleman (2009). "flowCore: a Bioconductor package for high throughput flow cytometry." BMC Bioinformatics **10**: 106.

Hamberger, B. and S. Bak (2013). "Plant P450s as versatile drivers for evolution of species-specific chemical diversity." Philos Trans R Soc Lond B Biol Sci **368**(1612): 20120426.

Heijne, G. (1986). "The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology." EMBO J **5**(11): 3021-3027.

Hessa, T., H. Kim, K. Bihlmaier, C. Lundin, J. Boekel, H. Andersson, I. Nilsson, S. H. White and G. von Heijne (2005). "Recognition of transmembrane helices by the endoplasmic reticulum translocon." Nature **433**(7024): 377-381.

Hessa, T., N. M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S. H. White and G. von Heijne (2007). "Molecular code for transmembrane-helix recognition by the Sec61 translocon." Nature **450**(7172): 1026-1030.

Hill, R. A. and J. D. Connolly (2015). "Triterpenoids." Nat Prod Rep **32**(2): 273-327.

Hofer, R., L. Dong, F. Andre, J. F. Ginglinger, R. Lugan, C. Gavira, S. Grec, G. Lang, J. Memelink, S. Van der Krol, H. Bouwmeester and D. Werck-Reichhart (2013). "Geraniol hydroxylase and hydroxygeraniol oxidase activities of the CYP76 family of cytochrome P450 enzymes and potential for engineering the early steps of the (seco)iridoid pathway." Metab Eng **20**: 221-232.

Holland, I. B. (2004). "Translocation of bacterial proteins--an overview." Biochim Biophys Acta **1694**(1-3): 5-16.

Holland, P. W. (2013). "Evolution of homeobox genes." Wiley Interdiscip Rev Dev Biol **2**(1): 31-45.

Hsu, M. F., T. F. Yu, C. C. Chou, H. Y. Fu, C. S. Yang and A. H. Wang (2013). "Using Haloarcula marismortui bacteriorhodopsin as a fusion tag for enhancing and visible expression of integral membrane proteins in Escherichia coli." PLoS One **8**(2): e56363.

Hull, A. K., R. Vij and J. L. Celenza (2000). "Arabidopsis cytochrome P450s that catalyze the first step of tryptophan-dependent indole-3-acetic acid biosynthesis." Proc Natl Acad Sci U S A **97**(5): 2379-2384.

Jensen, N. B., M. Zagrobelny, K. Hjerno, C. E. Olsen, J. Houghton-Larsen, J. Borch, B. L. Moller and S. Bak (2011). "Convergent evolution in biosynthesis of cyanogenic defence compounds in plants and insects." Nat Commun **2**: 273.

Jeong, K. J., M. J. Seo, B. L. Iverson and G. Georgiou (2007). "APEx 2-hybrid, a quantitative protein-protein interaction assay for antibody discovery and engineering." Proc Natl Acad Sci U S A **104**(20): 8247-8252.

Jorgensen, K., A. V. Rasmussen, M. Morant, A. H. Nielsen, N. Bjarnholt, M. Zagrobelny, S. Bak and B. L. Moller (2005). "Metabolon formation and metabolic channeling in the biosynthesis of plant natural products." Curr Opin Plant Biol **8**(3): 280-291.

Jorgensen, L., S. J. McKerrall, C. A. Kuttruff, F. Ungeheuer, J. Felding and P. S. Baran (2013). "14-step synthesis of (+)-ingenol from (+)-3-carene." Science **341**(6148): 878-882.

Kahn, R. A., T. Fahrendorf, B. A. Halkier and B. L. Moller (1999). "Substrate specificity of the cytochrome P450 enzymes CYP79A1 and CYP71E1 involved in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench." Arch Biochem Biophys **363**(1): 9-18.

Kaspera, R. and R. Croteau (2006). "Cytochrome P450 oxygenases of Taxol biosynthesis." Phytochem Rev **5**(2-3): 433-444.

Kell, D. B., N. Swainston, P. Pir and S. G. Oliver (2015). "Membrane transporter engineering in industrial biotechnology and whole cell biocatalysis." Trends Biotechnol **33**(4): 237-246.

King, A. J., G. D. Brown, A. D. Gilday, T. R. Larson and I. A. Graham (2014). "Production of bioactive diterpenoids in the euphorbiaceae depends on evolutionarily conserved gene clusters." Plant Cell **26**(8): 3286-3298.

Klepsch, M. M., J. O. Persson and J. W. de Gier (2011). "Consequences of the overexpression of a eukaryotic membrane protein, the human KDEL receptor, in Escherichia coli." J Mol Biol **407**(4): 532-542.

Kudla, G., A. W. Murray, D. Tollervey and J. B. Plotkin (2009). "Coding-sequence determinants of gene expression in Escherichia coli." Science **324**(5924): 255-258.

Kwon, S. K., S. K. Kim, D. H. Lee and J. F. Kim (2015). "Comparative genomics and experimental evolution of Escherichia coli BL21(DE3) strains reveal the landscape of toxicity escape from membrane protein overproduction." Sci Rep **5**: 16076.

Laursen, T., K. Jensen and B. L. Moller (2011). "Conformational changes of the NADPH-dependent cytochrome P450 reductase in the course of electron transfer to cytochromes P450." Biochim Biophys Acta **1814**(1): 132-138.

Lee, C., H. J. Kang, C. von Ballmoos, S. Newstead, P. Uzdavinys, D. L. Dotson, S. Iwata, O. Beckstein, A. D. Cameron and D. Drew (2013). "A two-domain elevator mechanism for sodium/proton antiport." Nature **501**(7468): 573-577.

Lee, J., N. Velmurugan and K. Jeong (2013). "Novel strategy for production of aggregation-prone proteins and lytic enzymes in Escherichia coli based on an anchored periplasmic expression system." Journal of bioscience and bioengineering.

Lee, J. W., D. Na, J. M. Park, J. Lee, S. Choi and S. Y. Lee (2012). "Systems metabolic engineering of microorganisms for natural and non-natural chemicals." Nat Chem Biol **8**(6): 536-546.

Leonard, E. and M. A. Koffas (2007). "Engineering of artificial plant cytochrome P450 enzymes for synthesis of isoflavones by Escherichia coli." Appl Environ Microbiol **73**(22): 7246-7251.

Lhor, M. and C. Salesse (2014). "Retinol dehydrogenases: membrane-bound enzymes for the visual function." Biochem Cell Biol **92**(6): 510-523.

Luirink, J., Z. Yu, S. Wagner and J. W. de Gier (2012). "Biogenesis of inner membrane proteins in Escherichia coli." Biochim Biophys Acta **1817**(6): 965-976.

Mao, C., C. E. Cheadle, S. J. Hardy, A. A. Lilly, Y. Suo, R. R. Sanganna Gari, G. M. King and L. L. Randall (2013). "Stoichiometry of SecYEG in the active translocase of Escherichia coli varies with precursor species." Proc Natl Acad Sci U S A **110**(29): 11815-11820.

Mazzoli, R. (2012). "Development of microorganisms for cellulose-biofuel consolidated bioprocessings: metabolic engineers' tricks." Comput Struct Biotechnol J **3**: e201210007.

Menzella, H. G. (2011). "Comparison of two codon optimization strategies to enhance recombinant protein production in Escherichia coli." Microb Cell Fact **10**: 15.

Mergulhao, F. J., D. K. Summers and G. A. Monteiro (2005). "Recombinant protein secretion in Escherichia coli." Biotechnol Adv **23**(3): 177-202.

Meunier, B., S. P. de Visser and S. Shaik (2004). "Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes." Chem Rev **104**(9): 3947-3980.

Mikkelsen, M. D., C. H. Hansen, U. Wittstock and B. A. Halkier (2000). "Cytochrome P450 CYP79B2 from Arabidopsis catalyzes the conversion of tryptophan to indole-3-acetaldoxime, a precursor of indole glucosinolates and indole-3-acetic acid." J Biol Chem **275**(43): 33712-33717.

Miroux, B. and J. E. Walker (1996). "Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels." J Mol Biol **260**(3): 289-298.

Mirzadeh, K., V. Martinez, S. Toddo, S. Guntur, M. J. Herrgard, A. Elofsson, M. H. Norholm and D. O. Daley (2015). "Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector-Coding Sequence Junction." ACS Synth Biol **4**(9): 959-965.

Mirzadeh, K., V. Martinez, S. Toddo, S. Guntur, M. J. Herrgard, A. Elofsson, M. H. Norholm and D. O. Daley (2015). "Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector-Coding Sequence Junction." ACS Synth Biol.

Mizutani, M. and D. Ohta (2010). "Diversification of P450 genes during land plant evolution." Annu Rev Plant Biol **61**: 291-315.

Moller, B. L. (2010). "Functional diversifications of cyanogenic glucosides." Curr Opin Plant Biol **13**(3): 338-347.

Monier, S., P. Van Luc, G. Kreibich, D. D. Sabatini and M. Adesnik (1988). "Signals for the incorporation and orientation of cytochrome P450 in the endoplasmic reticulum membrane." J Cell Biol **107**(2): 457-470.

Muller, D. J. and A. Engel (2007). "Atomic force microscopy and spectroscopy of native membrane proteins." Nat Protoc **2**(9): 2191-2197.

Murinova, S. and K. Dercova (2014). "Response mechanisms of bacterial degraders to environmental contaminants on the level of cell walls and cytoplasmic membrane." Int J Microbiol **2014**: 873081.

Murphy, C. D. (2012). "The microbial cell factory." Org Biomol Chem **10**(10): 1949-1957.

Nambara, E. and A. Marion-Poll (2005). "Abscisic acid biosynthesis and catabolism." Annu Rev Plant Biol **56**: 165-185.

Naray-Szabo, G., J. Olah and B. Kramos (2013). "Quantum mechanical modeling: a tool for the understanding of enzyme reactions." Biomolecules **3**(3): 662-702.

Nelson, D. R. (2013). "A world of cytochrome P450s." Philos Trans R Soc Lond B Biol Sci **368**(1612): 20120430.

Nelson, D. R., J. V. Goldstone and J. J. Stegeman (2013). "The cytochrome P450 genesis locus: the origin and evolution of animal cytochrome P450s." Philos Trans R Soc Lond B Biol Sci **368**(1612): 20120474.

Nielsen, A. Z., B. Ziersen, K. Jensen, L. M. Lassen, C. E. Olsen, B. L. Moller and P. E. Jensen (2013). "Redirecting photosynthetic reducing power toward bioactive natural product synthesis." ACS Synth Biol **2**(6): 308-315.

Nielsen, H., J. Engelbrecht, S. Brunak and G. von Heijne (1997). "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." Protein Eng **10**(1): 1-6.

Nielsen, J., M. Fussenegger, J. Keasling, S. Y. Lee, J. C. Liao, K. Prather and B. Palsson (2014). "Engineering synergy in biotechnology." Nat Chem Biol **10**(5): 319-322.

Nielsen, J. S. and B. L. Moller (2000). "Cloning and expression of cytochrome P450 enzymes catalyzing the conversion of tyrosine to p-hydroxyphenylacetaldoxime in the biosynthesis of cyanogenic glucosides in Triglochin maritima." Plant Physiol **122**(4): 1311-1321.

Norholm, M. H., F. Cunningham, C. M. Deber and G. von Heijne (2011). "Converting a marginally hydrophobic soluble protein into a membrane protein." J Mol Biol **407**(1): 171-179.

Norholm, M. H., S. Toddo, M. T. Virkki, S. Light, G. von Heijne and D. O. Daley (2013). "Improved production of membrane proteins in Escherichia coli by selective codon substitutions." FEBS Lett **587**(15): 2352-2358.

Paddon, C. J. and J. D. Keasling (2014). "Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development." Nat Rev Microbiol **12**(5): 355-367.

Park, H. G., Y. R. Lim, S. Han and D. Kim (2014). "Expression and Characterization of Truncated Recombinant Human Cytochrome P450 2J2." Toxicol Res **30**(1): 33-38.

Park, Y. and V. Helms (2008). "MINS2: revisiting the molecular code for transmembrane-helix recognition by the Sec61 translocon." Bioinformatics **24**(16): 1819-1820.

Petersen, T. N., S. Brunak, G. von Heijne and H. Nielsen (2011). "SignalP 4.0: discriminating signal peptides from transmembrane regions." Nat Methods **8**(10): 785-786.

Pires, N. D. and L. Dolan (2012). "Morphological evolution in land plants: new designs with old genes." Philos Trans R Soc Lond B Biol Sci **367**(1588): 508-518.

Podust, L. and D. Sherman (2012). "Diversity of P450 enzymes in the biosynthesis of natural products." Natural product reports **29**(10): 1251-1266.

Poulos, T. L. (2003). "The past and present of P450cam structural biology." Biochem Biophys Res Commun **312**(1): 35-39.

Pritchard, M., R. Ossetian, D. Li, C. Henderson, B. Burchell, C. Wolf and T. Friedberg (1997). "A general strategy for the expression of recombinant human cytochrome P450s in Escherichia coli using bacterial signal peptides: expression of CYP3A4, CYP2A6, and CYP2E1." Archives of biochemistry and biophysics **345**(2): 342-354.

Pritchard, M. P., R. Ossetian, D. N. Li, C. J. Henderson, B. Burchell, C. R. Wolf and T. Friedberg (1997). "A general strategy for the expression of recombinant human cytochrome P450s in Escherichia coli using bacterial signal peptides: expression of CYP3A4, CYP2A6, and CYP2E1." Arch Biochem Biophys **345**(2): 342-354.

Proschel, M., R. Detsch, A. R. Boccaccini and U. Sonnewald (2015). "Engineering of Metabolic Pathways by Artificial Enzyme Channels." Front Bioeng Biotechnol **3**: 168.

Quehl, P., J. Hollender, J. Schuurmann, T. Brossette, R. Maas and J. Jose (2016). "Co-expression of active human cytochrome P450 1A2 and cytochrome P450 reductase on the cell surface of Escherichia coli." Microb Cell Fact **15**: 26.

Ravichandran, K. G., S. S. Boddupalli, C. A. Hasermann, J. A. Peterson and J. Deisenhofer (1993). "Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's." Science **261**(5122): 731-736.

Rosano, G. L. and E. A. Ceccarelli (2014). "Recombinant protein expression in Escherichia coli: advances and challenges." Front Microbiol **5**: 172.

Samuelson, J. C., M. Chen, F. Jiang, I. Moller, M. Wiedmann, A. Kuhn, G. J. Phillips and R. E. Dalbey (2000). "YidC mediates membrane protein insertion in bacteria." Nature **406**(6796): 637-641.

Sarkar, D., N. Le Meur and R. Gentleman (2008). "Using flowViz to visualize flow cytometry data." Bioinformatics **24**(6): 878-879.

Schierle, C. F., M. Berkmen, D. Huber, C. Kumamoto, D. Boyd and J. Beckwith (2003). "The DsbA signal sequence directs efficient, cotranslational export of passenger proteins to the Escherichia coli periplasm via the signal recognition particle pathway." J Bacteriol **185**(19): 5706-5713.

Schlegel, S., M. Klepsch, D. Gialama, D. Wickstrom, D. J. Slotboom and J. W. de Gier (2010). "Revolutionizing membrane protein overexpression in bacteria." Microb Biotechnol **3**(4): 403-411.

Schlegel, S., E. Rujas, A. J. Ytterberg, R. A. Zubarev, J. Luirink and J. W. de Gier (2013). "Optimizing heterologous protein production in the periplasm of E. coli by regulating gene expression levels." Microb Cell Fact **12**: 24.

Seddon, A. M., P. Curnow and P. J. Booth (2004). "Membrane proteins, lipids and detergents: not just a soap opera." Biochim Biophys Acta **1666**(1-2): 105-117.

Sezutsu, H., G. Le Goff and R. Feyereisen (2013). "Origins of P450 diversity." Philos Trans R Soc Lond B Biol Sci **368**(1612): 20120428.

Shebley, M., U. M. Kent, D. P. Ballou and P. F. Hollenberg (2009). "Mechanistic analysis of the inactivation of cytochrome P450 2B6 by phencyclidine: effects on substrate binding, electron transfer, and uncoupling." Drug Metab Dispos **37**(4): 745-752.

Sibbesen, O., B. Koch, B. A. Halkier and B. L. Moller (1995). "Cytochrome P-450TYR is a multifunctional heme-thiolate enzyme catalyzing the conversion of L-tyrosine to p-hydroxyphenylacetaldehyde oxime in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench." J Biol Chem **270**(8): 3506-3511.

Singer, S. J. and G. L. Nicolson (1972). "The fluid mosaic model of the structure of cell membranes." Science **175**(4023): 720-731.

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

Sletta, H., A. Tondervik, S. Hakvag, T. E. Aune, A. Nedal, R. Aune, G. Evensen, S. Valla, T. E. Ellingsen and T. Brautaset (2007). "The presence of N-terminal secretion signal sequences leads to strong stimulation of the total expression levels of three tested medically important proteins during high-cell-density cultivations of Escherichia coli." Appl Environ Microbiol **73**(3): 906-912.

Smith, B. D., J. L. Sanders, P. R. Porubsky, G. H. Lushington, C. D. Stout and E. E. Scott (2007). "Structure of the human lung cytochrome P450 2A13." J Biol Chem **282**(23): 17306-17313.

Stephanopoulos, G. (2012). "Synthetic biology and metabolic engineering." ACS Synth Biol **1**(11): 514-525.

Studier, F. W. and B. A. Moffatt (1986). "Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes." J Mol Biol **189**(1): 113-130.

Saaf, A., M. Monne, J. W. de Gier and G. von Heijne (1998). "Membrane topology of the 60-kDa Oxa1p homologue from Escherichia coli." J Biol Chem **273**(46): 30415-30418.

Terpe, K. (2006). "Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems." Appl Microbiol Biotechnol **72**(2): 211-222.

Tsirigos, K. D., C. Peters, N. Shu, L. Kall and A. Elofsson (2015). "The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides." Nucleic Acids Res **43**(W1): W401-407.

Valent, Q. A., P. A. Scotti, S. High, J. W. de Gier, G. von Heijne, G. Lentzen, W. Wintermeyer, B. Oudega and J. Luirink (1998). "The Escherichia coli SRP and SecB targeting pathways converge at the translocon." EMBO J **17**(9): 2504-2512.

Vinothkumar, K. R. (2015). "Membrane protein structures without crystals, by single particle electron cryomicroscopy." Curr Opin Struct Biol **33**: 103-114.

von Heijne, G. (2006). "Membrane-protein topology." Nat Rev Mol Cell Biol **7**(12): 909-918.

Wagner, S., M. M. Klepsch, S. Schlegel, A. Appel, R. Draheim, M. Tarry, M. Hogbom, K. J. van Wijk, D. J. Slotboom, J. O. Persson and J. W. de Gier (2008). "Tuning Escherichia coli for membrane protein overexpression." Proc Natl Acad Sci U S A **105**(38): 14371-14376.

Wegerer, A., T. Sun and J. Altenbuchner (2008). "Optimization of an E. coli L-rhamnose-inducible expression vector: test of various genetic module combinations." BMC Biotechnol **8**: 2.

Yoon, S. H., M. J. Han, H. Jeong, C. H. Lee, X. X. Xia, D. H. Lee, J. H. Shim, S. Y. Lee, T. K. Oh and J. F. Kim (2012). "Comparative multi-omics systems analysis of Escherichia coli strains B and K-12." Genome Biol **13**(5): R37.

Zelasko, S., A. Palaria and A. Das (2013). "Optimizations to achieve high-level expression of cytochrome P450 proteins using Escherichia coli expression systems." Protein Expr Purif **92**(1): 77-87.

Zhang, J., Z. Kang, J. Chen and G. Du (2015). "Optimization of the heme biosynthesis pathway for the production of 5-aminolevulinic acid in Escherichia coli." Sci Rep **5**: 8584.

Zhang, X. and F. W. Studier (1997). "Mechanism of inhibition of bacteriophage T7 RNA polymerase by T7 lysozyme." J Mol Biol **269**(1): 10-27.

Zhao, S., R. Kumar, A. Sakai, M. W. Vetting, B. M. Wood, S. Brown, J. B. Bonanno, B. S. Hillerich, R. D. Seidel, P. C. Babbitt, S. C. Almo, J. V. Sweedler, J. A. Gerlt, J. E. Cronan and M. P. Jacobson (2013). "Discovery of new enzymes and metabolic pathways by using structure and genome context." Nature **502**(7473): 698-702.