

Vorhersage struktureller und biochemischer Eigenschaften von Cytochrom P450-Monooxygenasen und Laccasen

Von der Fakultät Energie-, Verfahrens- und Biotechnik der
Universität Stuttgart zur Erlangung der Würde eines Doktors der
Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von
DEMETGÜL SIRIM
aus Ludwigsburg

Hauptberichter: Prof. Dr. Jürgen Pleiss
Mitberichter: Prof. Dr. Bernhard Hauer

Tag der mündlichen Prüfung: 21.04.2010

Institut für Technische Biochemie der
Universität Stuttgart

2010

Cover

Kristallstruktur der Laccase aus *T. versicolor* (PDB-Eintrag 1GYC). Die in der Proteinfamiliendatenbank *LccED* annotierten charakteristischen Schleifen L1, M2, L3 und M4, die die Kupferbindestellen enthalten, sind verschieden farblich hervorgehoben. Die Kupfer-Ionen sind in orange dargestellt.

Teile dieser Arbeit wurden bereits veröffentlicht:

Fischer, M., Knoll, M., Sirim, D., Wagner, F., Funke, S., Pleiss, J., 2007. The Cytochrome P450 Engineering Database: A Navigation and Prediction Tool for the Cytochrome P450 Protein Family. *Bioinformatics* **23**: 2015-2017.

Copyright © 2007 Oxford University Press. Reprinted with kind permission

Sirim, D., Wagner, F., Lisitsa, A., Pleiss, J., 2009. The Cytochrome P450 Engineering Database: Integration of Biochemical Properties. *BMC Biochemistry* **10**: 27.

Weber, E., Sirim, D., Schreiber, T., Thomas, B., Pleiss, J., Hunger, M., Gläser, R., Urlacher, V.B., 2010. Immobilization of P450 BM-3 on mesoporous molecular sieves. *Journal of Molecular Catalysis B - Enzymatic* **64**: 29-37.

Copyright © 2010 Elsevier B.V. Reprinted with kind permission

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt entnommenen Gedanken sind als solche kenntlich gemacht.

Stuttgart, Januar 2010

(Demetgül Sirim)



Danksagungen

Ich danke Herrn Prof. Dr. Bernhard Hauer und Herrn Prof. Dr. Rolf D. Schmid für die Möglichkeit, die vorliegende Arbeit an ihrem Institut anfertigen zu können, sowie für die hervorragenden Arbeitsbedingungen und Unterstützung in allen Stadien der Doktorarbeit.

Mein besonderer Dank gilt Herrn Prof. Dr. Jürgen Pleiss für die Vergabe des interessanten Themas und für die ausgezeichnete wissenschaftliche Betreuung. Außerdem bedanke ich mich für die stetige Diskussionsbereitschaft und die Freiheit, die mir bei der Gestaltung der Arbeit blieb.

Herrn Prof. Dr. Michael Resch danke ich für die Bereitschaft, den Prüfungsvorsitz zu übernehmen.

Der Deutschen Forschungsgemeinschaft danke ich für die Finanzierung meiner Arbeit im Rahmen des Sonderforschungsbereiches 706. Den Kooperationspartnern, insbesondere Frau Prof. Dr. Vlada Urlacher, Evelyne Weber und Anna Romankiewicz, danke ich für die erfolgreiche Zusammenarbeit innerhalb des SFB706 und für die freundliche Beantwortung vieler Fragen.

Allen Mitarbeitern des Instituts danke ich für die offene und freundliche Zusammenarbeit und das gute Arbeitsklima. Dabei möchte ich besonders Christine Klug für ihre unermüdlige Hilfsbereitschaft danken. Sven Richter danke ich für viele interessante Gespräche und Motivation während der Zeit im *"front-office"*.

Meinen ehemaligen und derzeitigen Kollegen der Arbeitsgruppe "Bioinformatik" danke ich für die schöne Zeit, für die zahlreichen fruchtbaren Diskussionen und für die vielen gemeinsamen Unternehmungen. Besonders danken möchte ich Sascha Rehm, Michael Widmann, Alexander Steudle und Peter Trodler.

Den wissenschaftlichen Hilfskräften, die mich während meiner Promotion tatkräftig unterstützt haben, möchte ich sehr herzlich für ständige Bereitschaft hierfür danken. Dabei danke ich ganz besonders Florian Wagner, Thomas Hamm und Silvia Racolta.

Irina gilt mein besonderer Dank für ihre einzigartige Freundschaft und dafür, dass sie seit über 20 Jahren in jeder Situation für mich da war und mich immer weiter motiviert hat. Auch bei meinen Freunden Sanne, Ayca, Lena, Stephi, Peter und Bonzo möchte ich mich für die Unterstützung in den letzten Monaten bedanken.

Schließlich danke ich noch ganz besonders meinen Eltern, meiner Schwester Pinar und meinem Bruder Mehmet, meinem Schwager Lutz und meiner Schwägerin Antje, dafür, dass sie mir immer wieder zeigen, wie wichtig Zusammenhalt ist. Ohne ihre ständige Unterstützung, die weit über das Finanzielle hinausging, und ihre grenzenlose Geduld wäre diese Arbeit in dieser Form nicht möglich gewesen. Meinen Nichten Aliya, Luna, Melisa, Malika und Leana danke ich für das große Glück, das sie uns allen bescheren.

Für meine Eltern und Geschwister

Inhaltsverzeichnis

Erklärung	3
Danksagungen	5
Abbildungsverzeichnis	11
Abkürzungsverzeichnis	15
1 Zusammenfassung	19
2 Abstract	23
3 Einleitung	27
3.1 Cytochrom P450-Monooxygenasen	27
3.1.1 Reaktionsmechanismus	27
3.1.2 Redox-Partner	29
3.1.3 Nomenklatur	30
3.1.4 Strukturen und Sequenzdiversität	31
3.2 Laccasen und Multikupferoxidasen	33
3.2.1 Strukturen und Sequenzdiversität der Laccasen	34
3.3 Systematische Analyse von Proteinfamilien	36
3.3.1 Integration biologischer Daten	37
3.3.2 Relationale Datenbanken	38
3.3.3 Data Warehouse System	39
3.3.4 Vergleichende Sequenzanalyse	40
3.3.5 Probabilistische Modellierung	41
3.3.6 Von der Sequenz zur Funktion	42
4 Ziele dieser Arbeit	43

5	Ergebnisse und Diskussion	45
5.1	Cytochrom P450-Monooxygenasen	45
5.1.1	Die Cytochrome P450 Engineering Database: Ein Orientierungs- und Vorhersagewerkzeug für die Cytochrom P450 Proteinfamilie .	45
5.1.2	Die Cytochrome P450 Engineering Database: Integration bioche- mischer Eigenschaften	47
5.1.3	Die modulare Struktur von Cytochrom P450-Monooxygenasen . .	49
5.1.4	Modellierung von P450 BM-3 für Immobilisierungsversuche	52
5.2	Laccasen	55
5.2.1	Die Laccase Engineering Database: Ein Klassifikations- und Ana- lysesystem für Laccasen und verwandte Multikupferoxidasen . . .	55
6	Publikationsmanuskripte und Publikationen in englischer Sprache	59
6.1	The Cytochrome P450 Engineering Database: A Navigation and Prediction Tool for the Cytochrome P450 Protein Family	60
6.2	The Cytochrome P450 Engineering Database: Integration of Biochemical Properties	63
6.2.1	Abstract	63
6.2.2	Background	64
6.2.3	Construction and content	65
6.2.4	Utility and discussion	67
6.2.5	Conclusions	68
6.2.6	Availability and requirements	68
6.2.7	List of abbreviations	68
6.2.8	Authors' contributions	69
6.2.9	Acknowledgements	69
6.2.10	References	69
6.3	The Modular Structure of Cytochrome P450 Monooxygenases	71
6.3.1	Abstract	71
6.3.2	Background	72
6.3.3	Data	74
6.3.4	Methods	76
6.3.5	Results	78
6.3.6	Discussion	85
6.3.7	Conclusion	87

6.3.8	Authors' contributions	88
6.3.9	Acknowledgements	88
6.3.10	Supplementary material	89
6.3.11	References	93
6.4	Immobilization of P450 BM-3 monooxygenase on mesoporous molecular sieves with different pore diameters	97
6.5	The Laccase Engineering Database: A Classification and Analysis System for Laccases and Related Multicopper Oxidases	106
6.5.1	Abstract	106
6.5.2	Background	107
6.5.3	Construction and content	108
6.5.4	Contents	109
6.5.5	Utility and discussion	110
6.5.6	Conclusion	116
6.5.7	Availability and requirements	116
6.5.8	Authors' contributions	116
6.5.9	Acknowledgements	116
6.5.10	References	116
7	Gesamtliteraturverzeichnis	121

Abbildungsverzeichnis

3.1.1 Katalysezyklus von Cytochrom P450-Monooxygenasen. Verändert nach Denisov et al. (2005).	28
3.1.2 Schematische Übersicht über die Sekundärstruktur der Cytochrom P450-Monooxygenasen. Verändert nach Peterson und Graham (Peterson und Graham, 1998).	31
3.1.3 CYP- <i>fold</i> am Beispiel der 3-dimensionalen Struktur von Cytochrom P450 BM-3 aus <i>Bacillus megaterium</i> (PDB-Eintrag: 1BU7) in <i>cartoon</i> Darstellung. Die Häm-Gruppe ist in rot dargestellt.	32
3.2.1 Die globale Proteinfaltung der Laccasen, veranschaulicht am Beispiel der Laccase aus <i>Trametes versicolor</i> (PDB-Eintrag: 1GYC). Domäne 1 ist in rosa dargestellt, Domäne 2 in blau, Domäne 3 in grün und die Kupferionen in braun.	34
3.2.2 Das aktive Zentrum der Laccasen, veranschaulicht am Beispiel der Laccase aus <i>Trametes versicolor</i> (PDB-Eintrag: 1GYC). Das T1-Zentrum wird von His458, His395, Cys453 gebunden, während die Histidine 64, 66, 109, 111, 398, 400 und 452 den T2/T3-Cluster binden.	35
3.3.1 Zentrale Komponente des <i>Data Warehouse</i> : Die <i>ETL-Layer</i>	39
5.1.1 Die <i>feature page</i> der <i>CYPED</i> zeigt für jedes Protein in der Datenbank die Sequenz, ihre vorhergesagte Sekundärstruktur und die entsprechende Verknüpfung zur <i>CPK</i> an.	49
5.1.2 Molekulare Dimensionen der Häm-Domäne von P450 BM-3. Die Regionen der Kristallstruktur (PDB: 1BU7) sind in grün dargestellt, die modellierten Teile in blau. Die Größe des modellierten Proteins beträgt 80x70x60 Å ³	52
5.1.3 Die berechnete Titrationskurve des erweiterten Strukturmodells von P450 BM-3 zeigt eine bei ansteigendem pH Wert zunehmend negativ werdende Gesamtladung des Proteins mit einem <i>pI</i> von 5,4.	53

5.1.4	Elektrostatische Ladungsverteilung bei pH 7 an der Oberfläche des Strukturmodells der dreidimensionalen Struktur der Häm-Domäne und der FMN-Domäne von P450 BM-3 aus <i>Bacillus megaterium</i> (basierend auf den PDB-Einträgen 1BVY und 1BU7), in <i>surface</i> -Darstellung. Die Ladungsverteilung zeigt in proximaler Region einen positiven <i>patch</i> auf der Häm-Domäne, während die Oberfläche an der distalen Seite, wo sich auch der Substrateingangskanal befindet, negativ geladen ist. Häm- und FMN-Domäne des Proteins sind in <i>cartoon</i> dargestellt.	54
6.2.1	<i>CYPED</i> - <i>CPK</i> integration pipeline. Identification and assignment algorithm of the <i>CYPED</i> proteins and the corresponding <i>CPK</i> entries. The steps of the algorithm involve a BLAST search of each of the <i>CYPED</i> entries against the <i>CPK</i> , a ranking of the hits by E-value and a final pairwise alignment of the original <i>CYPED</i> entry with the corresponding <i>CPK</i> -hits to obtain the percentage identity.	66
6.3.1	Conserved regions derived from STAMP alignment mapped on reference structure P450 BM-3 from <i>Bacillus megaterium</i> (PDB: 1BU7). The SCRs are highlighted in blue, whereas the variable regions are shown in green. .	78
6.3.2	Conserved regions derived from STAMP alignment in a topological overview.	79
6.3.3	BC-loop region (SRS1) of CYPs. (A) Comparison of the BC-loops of P450 BM-3 (1BU7) in beige, CYP2C9 (1OG2) in green, CYP154C1 (1GWI) in pink, CYP101D (2CPP) in yellow and CYP107A1 (1OXA) in blue. (B) BC-loop region in P450 BM-3 (1BU7) and the position 87 corresponding residue in all 31 structures.	80
6.3.4	Amino acid composition of predicted F87 corresponding positions in all 8614 <i>CYPED</i> proteins. Green bars correspond to the percentage of aliphatic residues and phenylalanine, grey bars to amino acids of small polar nature and blue bars to charged amino acids. '-' denotes a gap in the alignment at the corresponding position.	82
6.3.5	Sites interacting with potential redox partners. The CPR-type FMN/FAD (PDB: 3ES9 from <i>Rattus norvegicus</i>) is shown in yellow, the parts of the P450 domains are shown in grey (PDB: 1OG2 from <i>Homo sapiens</i> , CYP2C9) and green (PDB: 2CPP from <i>Pseudomonas putida</i> , CYP101D), respectively. (A) Comparison RIS1 (α J/J' region) of the human CYP2C9 and P450cam CYP101D. (B) Comparison of RIS2 (meander insertion) of the human CYP2C9 and P450cam CYP101D.	83

6.3.6 (A) Fraction of proteins for each RIS1 (α J/J' region) length. (B) Fraction of proteins for each RIS2 (meander insertion) length.	84
6.3.7 STAMP alignment output. Stretches of residues with STAMP score $S_c > 6.0$ imply regions of conserved functions and are marked in black boxes.	92
6.5.1 Copper binding residues of laccase from <i>T. versicolor</i> (PDB entry 1GYC (Piontek et al., 2002)). The copper centres are shown in orange, the residues that match the defined pattern L1, M2, L3, M4 are coloured in red, green, blue, and yellow respectively (visualization by PyMOL (Delano, 2002)).	111
6.5.2 Phylogenetic tree for the homologous family I1 (Bilirubin Oxidases). The chosen coloring option is "by kingdom". Entries of bacterial origin are shown in blue, fungal entries in red, plant proteins in green and non-specified entries are colored in black.	114

Abkürzungsverzeichnis

$\mathcal{O}(\text{nm})$	polynomielle Laufzeit in Landau Notation
BLAST	<i>Basic Local Alignment Search Tool</i>
CPK	<i>Cytochrome P450 Knowledgebase</i>
CPR	Cytochrom P450-Reduktase
CYP	Cytochrom P450-Monooxygenase
CYPED	<i>Cytochrom P450 Engineering Database</i>
DB	Datenbank
DBMS	Datenbankmanagementsystem
DBS	Datenbanksystem
DSSP	<i>Define Secondary Structure of Proteins</i>
DWARF	<i>Data Warehouse for Analyzing Protein Families</i>
EC	<i>Enzyme commission (numbers)</i>
ETL	<i>Extraction, transformation, load</i>
FAD	Flavin-Adenin-Dinukleotid
FK	<i>foreign key</i> , Fremdschlüssel
FMN	Flavin-Mononucleotid
HMM	<i>Hidden Markov</i> Modell
LccED	<i>Laccase Engineering Database</i>
MCO	Multikupferoxidase
MSA	Multisequenzalignment
NADH	Nicotinamid-Adenin-Dinukleotid
NADPH	Nicotinamid-Adenin-Dinukleotid-Phosphat
P450 BM-3	Cytochrome P450-Monooxygenase BM-3 aus <i>Bacillus megaterium</i>
pI	isoelektrischer Punkt
PDB	<i>Protein Data Bank</i>
PK	<i>primary key</i> , Primärschlüssel
RIS	<i>Reductase interaction site</i>
SCR	<i>Structurally conserved region</i>

SNP *Single nucleotide polymorphism*
SQL *Structured Query Language*
SRS *Substrate recognition site*
STAMP *Structural Alignment of Multiple Proteins*
VMD *Visual Molecular Dynamics*

1 Zusammenfassung

Die Anwendung von Biokatalysatoren in industriellen Prozessen gewann in den letzten Jahren zunehmend an Bedeutung (Schmid et al., 2001). Mitglieder der Proteinfamilien der Cytochrom P450-Monooxygenasen und der Laccasen sind vielversprechende Kandidaten für solche industriellen Anwendungen, da sie die Oxidation einer Vielzahl verschiedener Substrate durch Luftsauerstoff katalysieren (Burton, 2003). Der Einsatz dieser Enzyme wird aber oft durch unzureichende Aktivitäten, mangelnde Selektivitäten oder andere ungeeignete Enzymeigenschaften erschwert. Mittlerweile ist es jedoch möglich, Enzyme zu entwickeln, die so modifiziert sind, dass sie in chemischen Verfahren einsetzbar sind. Dies setzt ein tiefgehendes Verständnis der Beziehungen zwischen Sequenz, Struktur und Funktion der Proteine der beiden Familien voraus.

Um die umfassende Analyse der Sequenz-Struktur-Funktionsbeziehungen der Familie der Cytochrom P450-Monooxygenasen (CYPs) zu ermöglichen, wurde innerhalb des *Data Warehouse Systems DWARF* (Fischer et al., 2006), das speziell für die Untersuchung von Proteinfamilien angelegt wurde, eine Proteinfamiliendatenbank erstellt. Die *CYPED* (*Cytochrome P450 Engineering Database*, <http://www.cyped.uni-stuttgart.de>) bildete damit das erste integrierte Datenbanksystem, mit dem Sequenzen, Strukturen und Funktionen der gesamten Proteinfamilie systematisch verglichen werden konnten. Die Erstveröffentlichung der Datenbank enthielt fast 4000 Proteineinträge, von denen für 25 unterschiedliche Proteine Kristallstrukturen abgelegt waren. Die Proteine wurden nach der Klassifizierung von Nelson (Nelson, 2006) in Superfamilien und homologe Familien unterteilt. Über das Web-Interface wurden den Nutzern Multisequenzalignments, phylogenetische Bäume und familienspezifische *HMM*-Profile für alle Familien zur Verfügung gestellt. Funktionell relevante Aminosäuren wurden in den Multisequenzalignments annotiert. Die *BLAST*-Suchfunktion innerhalb der Datenbank dient der Zuordnung neuer CYP-Proteine. Außer der erfolgreichen Anwendung dieser Datenbank für die Analyse von selektivitätsbestimmenden Aminosäuren wurde sie als Vorlage für den Aufbau weiterer Proteinfamiliendatenbanken und zum Abgleich der Klassen neuer CYP-Datenquellen verwendet.

Im Zuge einer Aktualisierung der *CYPED* wurde eine verfeinerte Klassifikation realisiert und die Datenbank neben neu hinzugekommenen Sequenz- und Strukturinformationen um neue Funktionalitäten erweitert, die eine noch umfassendere Analyse ermöglichen. Die neueste Version der Datenbank enthält fast die doppelte Datenmenge an Proteineinträgen. Mittels der Verlinkung der Einträge der *CYPED* mit den Einträgen der *CPK* (*Cytochrome P₄₅₀ Knowledgebase*, (Lisitsa et al., 2001)) konnten dem Nutzer nun auch Informationen über biochemische Eigenschaften, wie z.B. Substrate oder Inhibitoren der entsprechenden Proteine, zugänglich gemacht werden. Für die Verlinkung der Einträge wurde ein Algorithmus implementiert, bei dem die Verknüpfung über einen eindeutigen Primärschlüssel umgangen werden konnte. Dadurch konnten auch Proteineinträge verknüpft werden, die nicht über eine absolute Sequenzidentität verfügen. Des Weiteren wurde die *CYPED* um Informationen über Mutationen bzw. CYP-Polymorphismen erweitert. Von der *Home Page of Human Cytochrome P₄₅₀ (CYP) Allele Nomenclature Committee* (Oscarson und Ingelman-Sundberg, 2002) wurden dazu sämtliche Informationen über allelische Varianten und deren Effekte extrahiert und auf den Seiten der entsprechenden Proteineinträge aufgelistet. Da alle CYPs über eine hochkonservierte Faltung verfügen, war es möglich, aus einem Strukturalignment der verfügbaren CYP-Strukturen ein Profil abzuleiten, mit dem die konservierten Sekundärstrukturbereiche in allen Sequenzen innerhalb der *CYPED* annotiert werden konnten. Dies erlaubt die strukturelle Navigation in CYP-Sequenzen, für die keine Strukturinformationen vorhanden sind.

Die strukturierte Organisation der Sequenzinformationen von Cytochrom P450-Monooxygenasen ermöglichte den systematischen Vergleich der Sequenzen zur Identifikation von konservierten, für Struktur und Funktion essentiellen, Bereichen sowie von variablen Regionen, die für Selektivität und Interaktion mit Redox-Partnern verantwortlich sind. Es sind jedoch lediglich für etwa 5 % aller Proteine Strukturinformationen verfügbar, weshalb Rückschlüsse auf Funktionen und Reaktionsmechanismen sehr schwer durchführbar sind. Trotz der hohen Sequenzunterschiede verfügen alle CYPs über eine konservierte Faltung. Daher konnte aus einem Alignment, das aus der Überlagerung der Kristallstrukturen generiert wurde, ein verlässliches *HMM*-Strukturprofil entwickelt werden. Auf Basis dieses Profils bildet ein automatisiertes Skript jede Sequenz aus der *CYPED* auf das Alignment ab und kann anhand der Strukturinformationen konservierte Sekundärstrukturen in den Sequenzen vorhersagen. So konnten für alle *CYPED*-Sequenzen in einer automatischen Prozedur die konservierten Strukturbereiche vorhergesagt, annotiert und auf der aktuellen

online zugänglichen, Version der Datenbank veröffentlicht werden. Zusätzlich wurde das Skript in ein Web-Interface implementiert, mit dem für jede eingegebene CYP oder CYP-ähnliche Sequenz, die über die CYP-Faltung verfügt, eine Vorhersage erfolgen kann. Außerdem wurde durch die strukturelle Vorhersage die Analyse funktionell relevanter Positionen und Regionen, die sich in hochvariablen Sequenzregionen befinden, ermöglicht. So konnte die Position einer sich im aktiven Zentrum befindlichen Aminosäure, die aktivitäts-, regio- und stereoselektivitätsbestimmend zu sein scheint, in allen Sequenzen identifiziert und die Regionen, die mit potentiellen Redox-Partnern interagieren, analysiert werden.

Ein Nachteil bei der Verwendung von Cytochrom P450-Monooxygenasen als Biokatalysatoren ist ihre Instabilität. Die Immobilisierung auf Trägermaterialien kann sowohl ihre Stabilität als auch die Aktivität begünstigen. Daher wurden begleitend zu Immobilisierungs-Experimenten der Häm-Domäne von P450 BM-3 Berechnungen der biochemischen Eigenschaften des Proteins, wie pH-Abhängigkeit, elektrostatisches Potential und Struktureigenschaften, durchgeführt. Mit den Ergebnissen sollten dann Modelle zur Vorhersage der Interaktion mit dem Trägermaterial und der Orientierung des Proteins auf der Oberfläche erstellt werden. Durch die Modellierung der Struktur des in den Experimenten verwendeten Proteins konnten seine molekularen Dimensionen bestimmt werden und so ein Wert für die Porengröße der verwendeten Matrix abgeschätzt werden. Berechnungen der Titrationskurve ergaben einen pI von 5,4. Bei pH 7 wurde das optimale Immobilisierungsergebnis erzielt, bei einer negativen Gesamtladung des Proteins und der Matrix. Durch die Abbildung des elektrostatischen Potentials auf die Proteinoberfläche konnte gezeigt werden, dass die Bindung an die Matrix über einen positiven *patch* erfolgt, welcher bei pH 7 seine maximale Größe erreicht. Bei dieser Orientierung bindet die proximale Seite des Proteins, die normalerweise mit einer Reduktase interagiert, und ermöglicht so eine freie Substratpassage am distalen Zugangskanal.

Nach dem Vorbild der *CYPED* wurde als weitere Proteinfamiliendatenbank die *Laccase Engineering Database (LccED)*, <http://www.lcced.uni-stuttgart.de> erstellt. Sie enthält Sequenzdaten zu 2274 Laccasen und anderen homologen Multikupferoxidasen. Die Proteine wurden basierend auf phylogenetischen Analysen in Superfamilien und homologe Familien eingeteilt. Wie auch bei der *CYPED* wurden funktionell relevante Aminosäuren in den Multisequenzalignments annotiert. Die Familienverteilung wurde mit den Herkunftsorganismen der Proteinen abgeglichen. Außerdem wurden in der Literatur

beschriebene *pattern*, die die kupferbindenden Aminosäuren enthalten, in den Sequenzen identifiziert und ebenfalls annotiert. Durch die Validierung aller Sequenzen in der Datenbank konnte gezeigt werden, dass die *pattern*, welche auf einer geringen Sequenzanzahl basieren, zur Verallgemeinerung auf alle Sequenzen unzureichend waren. Daher wurden familien-spezifische *HMM*-Profile erstellt, die ebenso wie die Multisequenzalignments und phylogenetische Bäume über die Webseite zugänglich sind. Mittels dieser *HMM*-Profile können Proteine *in silico* klassifiziert und Laccasen von anderen Multikupferoxidasen differenziert werden. Für eine Zuordnung neuer Proteine zu den Klassen in der Datenbank steht die *BLAST*-Suchoption zur Verfügung.

Durch die systematische Analyse von Proteinfamilien sollten familienspezifische Eigenschaften untersucht und erklärt werden. Die Sammlung der relevanten Daten aus den verfügbaren *online*-Ressourcen in einem konsistenten, nichtredundanten Datenbanksystem ermöglichte erst eine solche Analyse in sinnvollem Umfang. Daher wurden für die beiden industriell relevanten Enzymklassen Cytochrom P450-Monooxygenasen und Laccasen Proteinfamiliendatenbanken etabliert. Basierend auf der sinnvollen Einteilung in Unterfamilien, umfangreichen Annotationen und systematischen Analysen dieser Datenbanken, die sowohl Sequenz und Struktur umfassten, konnten strukturelle und biochemische Eigenschaften beschrieben werden. Die erlangten Erkenntnisse wurden zur Vorhersage derselben Eigenschaften in noch nicht kristallisierten Proteinen verwendet. Des Weiteren konnten experimentelle Phänomene durch die einfache Modellierung biochemischer Eigenschaften beschrieben werden.

2 Abstract

In recent years, the use of biocatalysis in industrial processes is gaining more and more importance (Schmid et al., 2001). Members of the protein families of cytochrome P450 monooxygenases and laccases are interesting candidates for such applications, since they are able to catalyze the oxidation of a large variety of substrates by molecular oxygen (Burton, 2003). Still, their broader application is limited because of inadequate enzymatic properties and low activities. However, the modification of enzymes with the aim to render them more suitable for chemical applications is possible. The prerequisite for this purpose is a deep insight into the relationships between sequence, structure and function of the proteins in these families.

To allow a comprehensive analysis of the sequence-structure-function relationships within the vast and diverse family of cytochrome P450 monooxygenases (CYPs), a protein family database within the data warehouse system for the analysis of protein families *DWARF* (Fischer et al., 2006) was established. Therefore, the *CYPED* (*Cytochrome P450 Engineering Database*, <http://www.cyped.uni-stuttgart.de>) was the first integrated database system for the systematic comparison of sequences, structures and function of the complete CYP protein family. Its first release contained almost 4000 protein entries, for 25 of them crystal structures were deposited. The proteins were classified in superfamilies and homologous families according the Nelson classification scheme (Nelson, 2006). Multisequence alignments for the families, phylogenetic trees and family-specific HMM profiles were provided via a web interface. Functionally relevant amino acids were annotated within the multisequence alignments. For the classification of new CYP proteins, a *BLAST* search can be performed against the *CYPED*. The *CYPED* was successfully applied in the analysis of selectivity determining residues, served as template for the establishment of other protein family databases and was applied in the adjustment of the families in other CYP data sources.

In the following update procedure a new version with a refined classification was realized and, besides integrating new sequences and structures, the *CYPED* was extended by biochemical properties, and by adding new functionalities. The amount of sequences and structures almost doubled. The *CYPED* entries could be linked to the entries of the *CPK* (*Cytochrome P450 Knowledgebase*, (Lisitsa et al., 2001)) and therefore data on biochemical properties, like substrates or inhibitors, could be provided. For this purpose a new algorithm was implemented, where the linkage of the entries could be realized by their high sequence similarity instead of a unique primary key. Thus, it was possible to link sequences which belong to the same protein entries although not being absolutely identical. Furthermore, information on mutations and CYP polymorphism was added. All data on allelic variants and their effects were extracted from the *Home Page of Human Cytochrome P450 (CYP) Allele Nomenclature Committee* (Oscarson und Ingelman-Sundberg, 2002) and listed on the respective protein entries. Since all CYPs share a common conserved fold, it was possible to derive a profile based on a structural alignment. With this profile all highly conserved secondary structure elements could be predicted and annotated among the *CYPED* entries, hence allowing a structural navigation in sequences without structural information.

The structured organization of the sequence information on cytochrome P450 monooxygenases allows the systematic comparison of all sequences in order to identify conserved regions which are essential for structure and function and variable regions responsible for selectivity and interaction with redox partners. Unfortunately, structural information is limited to only 5 % of all CYPs. This makes it extremely difficult to draw conclusions on function and mechanism of the proteins. However, despite the profoundly low sequence similarity, the structures of all CYPs are preserved highly similar. Therefore, an alignment which was based on a superposition of all available CYP structures could be generated, which led to an HMM profile of conserved and variable regions. An automated script was developed which is able to map each *CYPED* sequence on the alignment and consequently predict secondary structures based on the provided structural information. Thus, it was possible to predict and annotate the conserved secondary structures among all *CYPED* sequences in an automated procedure. The results were published in the current online version of the database. Additionally, a web interface was implemented, which can be applied to perform such a prediction for any entered CYP sequences or similar sequences with the same fold. Furthermore, the structural prediction allowed the analysis of functionally relevant positions and sites in highly variable regions. Thus, a

key residue in determination activity, regio- and stereoselectivity could be identified in all sequences and the reductase interaction sites could be analyzed.

A limitation of the application of cytochrome P450 monooxygenases as biocatalysts is their inherent low stability. Their immobilization increases their stability and activity. Therefore, the heme domain of P450 BM-3 was immobilized on mesoporous materials with the aim to render the protein more suitable for synthetic applications. Supporting these experiments, biochemical properties of the protein, like pH-dependence, its electrostatic properties and structural characteristics could be revealed. The results should assist in the establishment of reliable models in order to predict the interaction with the matrix and the orientation of the protein on it. By modelling the exact structure of the protein which was immobilized in the experiments, its molecular dimensions could be calculated and the pore-size of the mesoporous material which was used could be estimated. The calculated titration curve showed a pI of 5.4, and, that at pH 7, where the most amount of protein could be adsorbed on the surface, the protein has a negative net-charge. By calculating the electrostatic potential and mapping it on the protein surface, a positive patch on the negatively charged surface could be detected. This patch reaches its maximum size at pH 7 and mediates the binding to the surface. In this binding-mode on the proximal site of the protein, which typically interacts with the reductase, the substrate entrance channel on the distal site is not inhibited and allows a free access to the active site.

Further, the protein family database approach was applied to the family of multicopper oxidases in order to facilitate the systematic analysis of this protein family. The resulting *Laccase Engineering Database* (*LccED*, <http://www.lcced.uni-stuttgart.de>) contains sequence data on 2274 laccases and their homologous multicopper oxidases. The proteins were classified in superfamilies and homologous families based on phylogenetic studies. The overall distribution of source organisms among the families was compared with the initial classification. As with the *CYPED*, functionally relevant amino acids were annotated among the sequences. Sequence patterns, described in literature, comprising the copper-binding residues were identified in the sequences and annotated as well. A manual validation of all sequences in the *LccED* led to the conclusion, that the patterns which were based on a low number of sequences could not be generalized for all laccase and multicopper oxidase sequences. Therefore, family-specific *HMM*-profiles were provided for each family, which are available via the web page, as well as multisequence alignments and phylogenetic trees. These *HMM*-profiles support the *in silico* classification of new

proteins and can be applied to separate laccases from other multicopper oxidases. An assignment of new multicopper proteins may also occur via a *BLAST* search against the LccED.

By means of a systematic analysis of proteins in the context of their entire family of homologous proteins, structural and biochemical properties should be explained. By collecting the relevant data from several online sources and storing it in a consistent, non-redundant database system such an analysis is possible within a reasonable extent. Therefore, protein family databases for the industrially relevant enzyme classes cytochrome P450 monooxygenases and laccases were established. By the systematic classification, consistent annotation and the comprehensive analysis of these databases, involving both sequence and structure, structural and biochemical properties could be explained and the newly gained insights could be transferred on proteins with no structural information. Modelling of biochemical properties contributed to the explanation of experimental observations.

3 Einleitung

Mittels spezifischer Mutationen durch rationales Design können bestimmte Eigenschaften in Enzymen optimiert werden. Dazu ist neben der Kenntnis der Aminosäureabfolge und der Struktur des Proteins auch ein Verständnis der Zusammenhänge zwischen Sequenz, Struktur und Funktion notwendig. Daher wurden *in silico*-Methoden angewandt, um die Sequenz-Struktur-Funktionsbeziehung der beiden Proteinklassen Cytochrom P450-Monooxygenasen und Laccasen, dahingehend zu erforschen. Vertreter beider Klassen sind von großem Interesse für biotechnologische Anwendungen, da sie die Oxidation einer Vielzahl verschiedener Substrate durch Luftsauerstoff katalysieren (Burton, 2003). Die folgenden Kapitel geben eine Einführung in die beiden Enzymklassen und in die theoretischen Grundlagen der angewandten bioinformatischen Methoden.

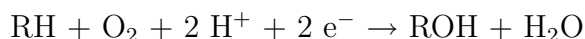
3.1 Cytochrom P450-Monooxygenasen

Cytochrom P450-Monooxygenasen (CYPs) bilden eine der größten Enzymfamilien in der Natur und werden der Klasse der Monooxygenasen zugeordnet (EC 1.14.-.-). Sie enthalten ein Häm-Porphyrin als prosthetische Gruppe, dessen Bindung über das Eisen an das Protein über ein Cystein erfolgt. Ihren Namen erhielten diese Proteine wegen ihres charakteristischen Absorptionsspektrums bei 450 nm im Kohlenmonooxiddifferenzspektrum (Klingenberg, 1958; Garfinkel, 1958). Sie kommen in fast allen eukaryotischen und prokaryotischen Lebensformen vor und besitzen Schlüsselfunktionen im Metabolismus körpereigener und körperfremder Verbindungen bei Säugetieren. Daher sind sie von großer Bedeutung in der pharmazeutischen Industrie (Guengerich, 1991). Außerdem katalysieren sie stereo- und regioselektiv eine Vielzahl chemischer Reaktionen und sind daher als vielseitige Biokatalysatoren von großer industrieller Relevanz (Urlacher und Eiben, 2006).

3.1.1 Reaktionsmechanismus

Die katalytische Vielfalt von Cytochrom P450-Monooxygenasen reicht von Hydroxylierungsreaktionen über N-, O- und S-Dealkylierung, Desaminierung, Entschwefelung und

Dehalogenierung bis zur Epoxidierung verschiedenster Substrate (Guengerich, 2001) nach folgender Reaktionsgleichung:



Dabei wird ein Sauerstoffatom in das Substrat eingeführt und das andere zu Wasser reduziert. Die zwei benötigten Elektronen werden von Redox-Partnern bereitgestellt. Die einzelnen Schritte des heute allgemein akzeptierten Reaktionszyklus (Denisov et al., 2005) ist in Abbildung 3.1.1 dargestellt.

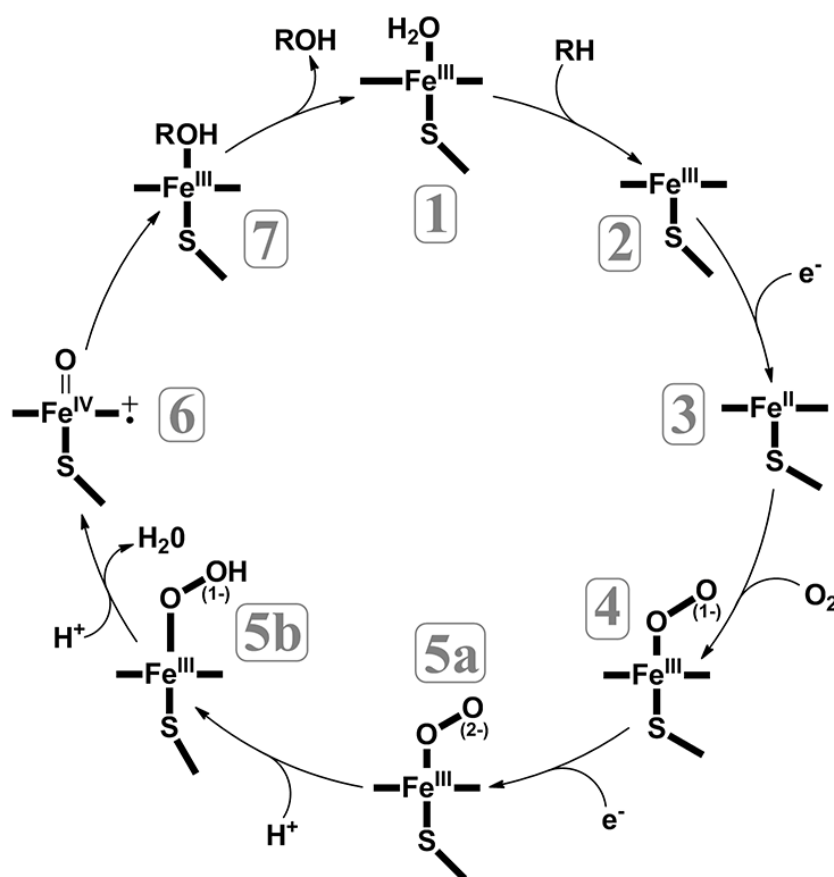


Abbildung 3.1.1: Katalysezyklus von Cytochrom P450-Monooxygenasen. Verändert nach Denisov et al. (2005).

Der Katalysezyklus startet in der Ruheform des Enzyms, wobei sich das Fe^{3+} -Zentrum in einem *low-spin*-Zustand befindet (1). Bei Substratbindung wird der Wasser-Ligand vom Eisenzentrum verdrängt und es kommt als Folge der Erniedrigung der Koordinationszahl

des Eisens zu einem *low-spin-/high-spin*-Übergang (2). Anschließend findet eine Reduktion des Fe^{3+} zu Fe^{2+} durch den Redox-Partner statt (3). Fe^{2+} ist in der Lage, molekularen Sauerstoff zu binden und bildet dann einen Oxy-Eisen-Komplex (4). Durch die erneute Reduktion dieses Komplexes erhält man ein Peroxo-Eisenintermediat (5a), welches durch Protonierung zum Hydroperoxo-Eisenintermediat umgesetzt wird. Durch eine weitere Protonierung und Wasserabspaltung erhält man den Oxo-Eisenkomplex (6). Als letzter Schritt erfolgt die Oxygenierung des Substrats durch einen radikalischen *rebound*-Mechanismus (7). Nach der Abspaltung des monooxygenierten Substrats stellt die Anlagerung eines Wasser-Liganden wieder die Ruheform her.

3.1.2 Redox-Partner

Zur Aktivierung des Luftsauerstoffs benötigen Cytochrom P450-Monooxygenasen zwei Elektronen, die in den meisten Fällen sequentiell von einem Redox-Partner zur Verfügung gestellt werden. Inzwischen sind eine Reihe verschiedener Elektronentransfersysteme bekannt (Hannemann et al., 2007; Munro et al., 2007). Die Klassifizierung der CYPs nach den vier wichtigsten Elektronentransfersystemen lautet nach McLean et al. (2005):

- **Klasse I:** Mitochondriale und die meisten bakteriellen CYPs, wie z. B. P450*cam* aus *Pseudomonas putida*, gehören zu dieser Klasse. Sie benötigen für den Elektronentransport eine NADPH- oder NADH-abhängige FAD-enthaltende Reduktase und ein 2Fe-2S-Ferredoxin-Enzym.
- **Klasse II:** Die mikrosomalen CYPs sind membrangebundene Zwei-Komponenten-Systeme, bestehend aus einer P450-Häm-Domäne sowie einer FAD und FMN enthaltenden Cytochrom P450-Reduktase (CPR).
- **Klasse III:** Bei Proteinen dieser Klasse, wie zum Beispiel P450 BM-3 aus *Bacillus megaterium*, einem der am häufigsten in der Biokatalyse angewandten Enzyme, ist die Häm-Domäne mit einer FMN/FAD-Reduktase-Domäne fusioniert.
- **Klasse IV:** Klasse-IV-CYPs, wie zum Beispiel in *Rhodococcus*, sind lösliche Ein-Komponenten-Enzyme, bestehend aus einer Häm-Domäne fusioniert mit einer FMN-Reduktase und einer Ferredoxin-Domäne.

Die Interaktion der CYPs mit ihren entsprechenden Elektronen-Donoren ist eine wichtige Voraussetzung für das Stattfinden der Reaktion. Da beispielsweise in den Zellen der Leber verschiedene CYPs mit ein und derselben Reduktase interagieren, wird

angenommen, dass sich die CYPs in ihren Reduktaseinteraktionsflächen stark ähneln. Dabei sollen Salzbrücken für die Interaktion und die Orientierung der Proteine zueinander verantwortlich sein (Bernhardt, 1996). Eine Struktur für Gesamtkomplexe von CYPs und ihren Redox-Partnern ist nicht verfügbar. Da der Elektronentransfer zu der P450-Häm-Domäne geschwindigkeitsbestimmend zu sein scheint (Guengerich, 2002), stellt das Verständnis der Interaktion der Komponenten der Elektronentransfersysteme eine große Herausforderung dar. Die Analyse der interaktionsbestimmenden Regionen der CYPs und das daraus erlangte Verständnis kann zum Design von Systemen mit optimierten Interaktionen und damit zu signifikant verbesserten Biokatalysatoren führen (Bernhardt, 2006).

3.1.3 Nomenklatur

Die Nomenklatur der CYPs erfolgt anhand von Sequenzhomologien auf Aminosäureebene. Sie werden eingeteilt in Genfamilien und Subfamilien, sowie jeweils in die entsprechenden Isoformen. Das seit 1989 etablierte systematische Nomenklaturschema (Nebert et al., 1989) basiert auf der 1987 vorgeschlagenen Nomenklatur (Nebert et al., 1987), bis 2006 wurden zudem zahlreiche Aktualisierungen veröffentlicht (Nelson, 2006). Nach diesem Schema galt ursprünglich, dass alle CYPs einer Genfamilie (Superfamilie) über 40 % Sequenzidentität verfügen und die CYPs einer Subfamilie (homologe Familie) in mindestens 55 % ihrer Aminosäuren übereinstimmen müssen. Bis auf einige Ausnahmen ist diese Aufteilung auch heute noch allgemein gültig.

Die Bezeichnung für ein CYP-Gen setzt sich zusammen aus dem Präfix CYP, einer Zahl für die Genfamilie, einem Großbuchstaben, der die Subfamilie bezeichnet und einer weiteren Zahl für das entsprechende Gen. So steht beispielweise die Abkürzung CYP102A1 für P450 BM-3 aus *Bacillus megaterium*, dem 1. Gen, das in der Subfamilie A der Familie 102 entdeckt wurde. Eine besondere Konvention gilt für die Bezeichnung der CYP-Gene aus Maus und Drosophila, für welche kursive Kleinbuchstaben verwendet werden (z. B. *Cyp1a1*).

Mit der rapide ansteigenden Zahl der bekannten CYP-Gene und der damit immer höher werdenden Diversität der Sequenzen untereinander wurden auch phylogenetische Betrachtungen zur Klassifikation von CYP-Genen erforderlich. Die Anzahl der bekannten CYP-Gene lag im August 2009 bei über 11000 (laut Nelson, <http://drnelson.utmem.edu/>). Für die einheitliche Nomenklatur ist das *Committee on Standardized Cytochrome P450 Nomenclature* unter Nelson zuständig.

3.1.4 Strukturen und Sequenzdiversität

Aufgrund der beträchtlichen Sequenzdiversität der Cytochrom P450-Monooxygenasen zeigen Proteine aus unterschiedlichen Familien Sequenzidentitäten von teilweise unter 15 %. Trotzdem verfügen alle CYPs über gemeinsame strukturelle Merkmale, die besonders im Bereich des aktiven Zentrums hoch konserviert sind. Bereits nach der Kristallisation der ersten Struktur P450*cam* aus *Pseudomonas putida* (CYP101) im Jahr 1987 wurde eine Bezeichnung der α -Helices (A-L) und der β -Faltblattstrukturen (1-4), sowie eine Unterteilung in zwei Domänen, einer α - und einer β -Domäne, vorgeschlagen (Poulos et al., 1987). Die bisher aufgelösten fast 40 Strukturen verschiedener CYP-Proteine bestätigen, dass bei allen CYPs der Häm-Porphyrin-Ring im katalytischen Zentrum von einem Vier-Helix-Bündel, bestehend aus den aneinanderliegenden α -Helices D, E, I und L, sowie

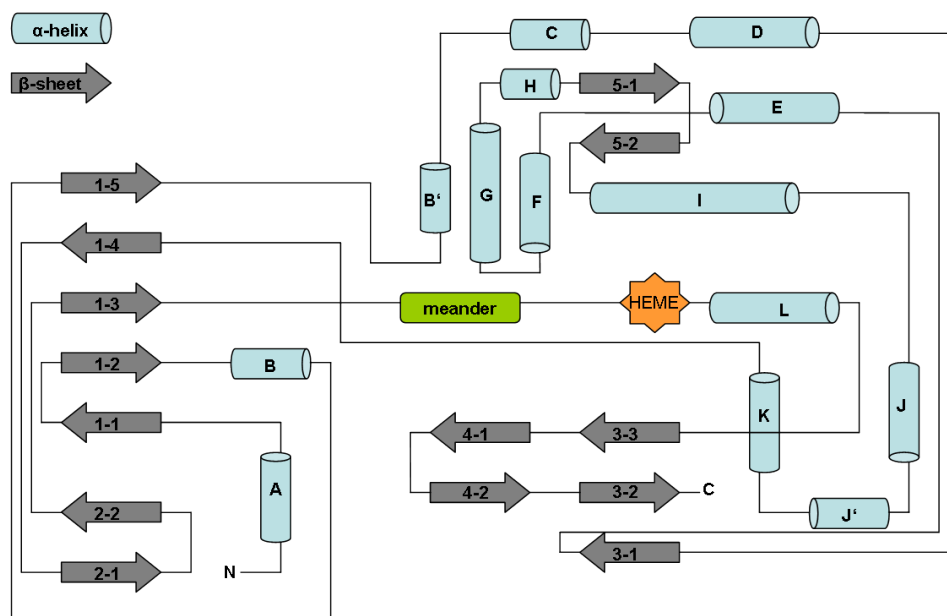


Abbildung 3.1.2: Schematische Übersicht über die Sekundärstruktur der Cytochrom P450-Monooxygenasen. Verändert nach Peterson und Graham (Peterson und Graham, 1998).

den Helices J und K, sowie zwei β -Faltblattstrukturen und der Häm-stabilisierenden Meander-Schleife (Hasemann et al., 1995; Peterson und Graham, 1998) umgeben ist (Abbildung 3.1.2).

Dieser strukturell konservierte Kern ist essentiell für Struktur und Funktion. Sequenz- und strukturumfassende Analysen zeigten außerdem, dass CYPs über variable Regionen verfügen, die für Substraterkennung und -bindung, sowie für die Interaktion mit den jeweiligen Redox-Partnern verantwortlich sind (Mestres, 2005). Auf der weniger konservierten, distalen Seite der Häm-Gruppe wurden sechs solcher Substratbinderegionen (SRS, *substrate recognition sites*) identifiziert und mit SRS1-SRS6 bezeichnet (Gotoh, 1992). Trotz der Abweichungen in diesen Bereichen wird der beschriebene CYP-*fold* jedoch in allen CYP-Strukturen eingehalten (Abbildung 3.1.3).

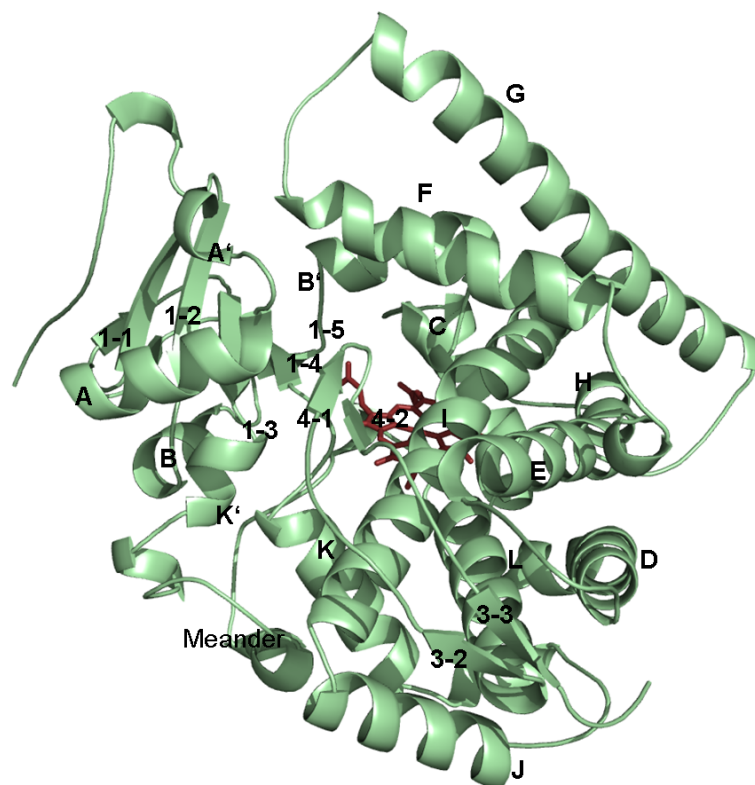


Abbildung 3.1.3: CYP-*fold* am Beispiel der 3-dimensionalen Struktur von Cytochrom P450 BM-3 aus *Bacillus megaterium* (PDB-Eintrag: 1BU7) in *cartoon* Darstellung. Die Häm-Gruppe ist in rot dargestellt.

Auf Sequenzebene absolut konserviert ist lediglich der proximale Cysteinligand des Häms, umgeben von einem hochkonservierten Konsensusmotiv. Weiterhin hochkonserviert sind

das in fast allen CYPs vorhandene EXXR-Motiv (Rupasinghe et al., 2006), das die Meander-Region stabilisiert und am Elektronentransport beteiligt zu sein scheint, und einem ebenfalls in den meisten CYPs zu findenden Threoninrest in der Helix I, der wahrscheinlich am Protonentransfer zum aktiven Zentrum beteiligt ist (Hasemann et al., 1995).

Um aus den Unterschieden und Gemeinsamkeiten verschiedener CYP-Familien Regeln für biochemische Eigenschaften für z. B. Spezifität, Selektivität oder auch Reduktase-Interaktionen abzuleiten, sind systematische Analysen der Proteine im Kontext der gesamten Proteinfamilie von großer Bedeutung. Verschiedene Quellen im Internet bieten CYP-Sequenzen und -Strukturen an (Nelson, 2002), teilweise auch mit Verweisen auf biochemische Eigenschaften (Lisitsa et al., 2001). Dennoch gab es bislang keine integrierte Sammlung der relevanten Daten, die in Kombination mit geeigneten Analysewerkzeugen eine solche Untersuchung familienspezifischer Eigenschaften möglich macht.

3.2 Laccasen und Multikupferoxidasen

Multikupferoxidasen (*multi copper oxidases*, MCOs), oder auch „blaue Multikupferproteine“ genannt, sind Enzyme, welche die Ein-Elektronen-Oxidation einer Vielzahl, vorwiegend phenolischer, Substrate bei gleichzeitiger Vier-Elektronen-Reduktion von molekularem Sauerstoff zu Wasser katalysieren (Chalupský et al., 2006):



Die Klasse der MCOs umfasst vier Enzymfamilien (Solomon et al., 1996):

- Laccasen (EC 1.10.3.2)
- Ascorbatoxidasen (EC 1.10.3.3)
- Ferroxidasen (EC 1.16.3.1)
- Caeruloplasmin, ebenfalls (EC 1.16.3.1)

Insbesondere Laccasen werden als ideale ”grüne“ Katalysatoren von hohem biotechnologischem Potential betrachtet, da sie trotz ihrer geringen Ansprüche – sie benötigen lediglich Luftsauerstoff und produzieren nur Wasser als Nebenprodukt – ein breites Substratspektrum besitzen und eine Vielzahl an Reaktionen katalysieren (Kunamneni et al., 2008). Ihre Anwendungen reichen von der Textil-, Papier- und Zellstoffindustrie,

über Lebensmittelindustrie bis hin zur biologischen Dekontaminierung und organischen Synthese (Riva, 2006). Die erste Laccase wurde bereits 1883 aus dem Pflanzensaft des japanischen Lackbaumes (*Rhus vernicifera*) isoliert und hat daher ihren Namen erhalten (Yoshida, 1883). Laccasen sind weit verbreitet und kommen neben Pflanzen auch in Pilzen, Bakterien und Pflanzen vor, wo sie unterschiedliche physiologische Funktionen einnehmen (Mayer und Staples, 2002). In Pilzen sind sie am Ligninabbau und an der Pigment- und Sporensynthese beteiligt, in Pflanzen katalysieren sie die Ligninbiosynthese und tragen zum Zellwandaufbau bei. Bei Bakterien sollen sie für den Aufbau UV-resistenter Sporen verantwortlich sein (Tadesse et al., 2008).

3.2.1 Strukturen und Sequenzdiversität der Laccasen

Aus den bekannten Multikupferoxidase-Kristallstrukturen für Ferroxidasen, Oxidasen, Cearuloplasmin und Laccasen wird ersichtlich, dass kupferhaltige Proteine aus mehreren Domänen bestehen, die als Cupredoxin-Domänen beschrieben wurden (Adman, 1991; Murphy et al., 1997). Die Analyse der verfügbaren Kristallstrukturen für Laccasen zeigt, dass sie aus drei solcher Cupredoxin-ähnlichen Domänen D1, D2 und D3 aufgebaut sind (Abbildung 3.2.1), ähnlich wie auch bei den Strukturen der Ascorbatoxidasen. Die Domänen weisen vorwiegend eine β -Faltblattstruktur auf.

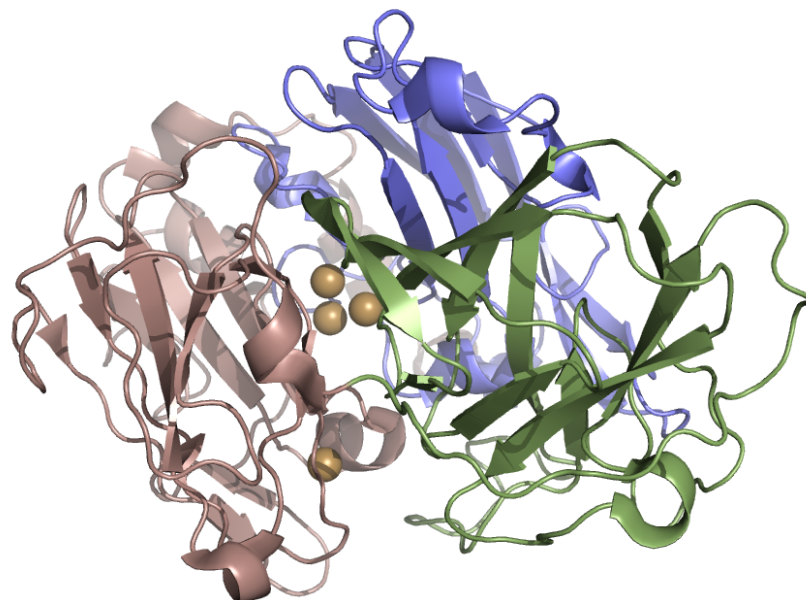


Abbildung 3.2.1: Die globale Proteinfaltung der Laccasen, veranschaulicht am Beispiel der Laccase aus *Trametes versicolor* (PDB-Eintrag: 1GYC). Domäne 1 ist in rosa dargestellt, Domäne 2 in blau, Domäne 3 in grün und die Kupferionen in braun.

Laccasen enthalten vier Kupferionen, die auf zwei aktive Zentren verteilt sind. Das T1-Kupfer-Zentrum, bestehend aus einem Typ1-Kupferion, befindet sich in der Domäne 3. Hier findet die Substratbindung und -oxidation statt. Das trinukleäre T2/T3-Zentrum, an dem Sauerstoff reduziert wird, besteht aus einem Typ2- und zwei Typ3-Kupferionen und liegt an der Schnittstelle zwischen D1 und D3 (Piontek et al., 2002; Bertrand et al., 2002). In der Laccase aus *Trametes versicolor* binden neun Histidine und ein Cystein die vier Kupferionen (Abbildung 3.2.2).

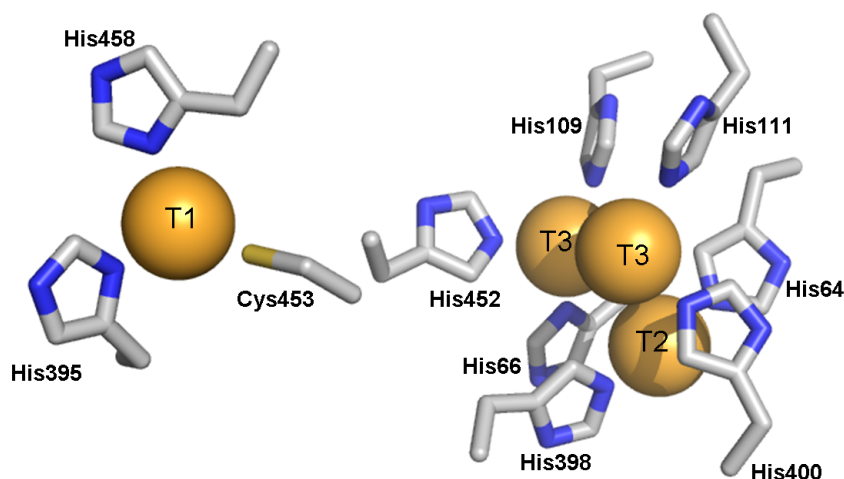


Abbildung 3.2.2: Das aktive Zentrum der Laccasen, veranschaulicht am Beispiel der Laccase aus *Trametes versicolor* (PDB-Eintrag: 1GYC). Das T1-Zentrum wird von His458, His395, Cys453 gebunden, während die Histidine 64, 66, 109, 111, 398, 400 und 452 den T2/T3-Cluster binden.

Phylogenetische Studien, die Sequenzen und Strukturen unterschiedlicher Cupredoxin-Proteine umfassen, deuteten auf eine funktionelle und molekulare Evolution der Multikupferoxidasen von Proteinen hin, die aus einer Cupredoxin-Domäne bestehen, bis hin zu Mehrdomänenproteinen (Nakamura und Go, 2005). Die kürzlich kristallisierte Zwei-Domänen-MCO aus *Nitrosomonas europaea* bestätigte diese Hypothese (Lawton et al., 2009).

Neben den strukturell konservierten Bereichen, die hauptsächlich in den für Struktur und Funktion essentiellen Regionen liegen, gibt es auch bei Multikupferoxidasen variable *loops*, die für die Substratbindung verantwortlich sein sollen. Diese *loops* wurden erstmals von Larrondo et al. (2003) beschrieben und bilden ein vierflügeliges Tor von unterschiedlicher Flexibilität. Durch den somit variablen Substratseingangskanal wird die unterschiedliche Substratakzeptanz und das Substratbindungsvermögen der verschiedenen Enzyme erklärt.

Auf Sequenzebene variieren die Sequenzen der Multikupferoxidasen und auch die der Laccasen untereinander stark. Eine phylogenetische Unterteilung basierend auf Sequenzähnlichkeit ergab zehn verschiedene verwandte Unterfamilien der Multikupferoxidasen: **(A)** Basidiomycete-Laccasen, **(B)** Ascomycete-Laccasen, **(C)** Insekten-Laccasen, **(D)** pigmentbildende MCOs aus Pilzen, **(E)** pilzliche Ferroxidasen, **(F)** pilzliche und pflanzliche Ascorbatoxidasen, **(G)** pflanzliche laccase-ähnliche MCOs, **(H)** Kupfer-Resistenzproteine (CopA), **(I)** Bilirubinoxidasen, und **(J)** Proteine des Kupferflußsystems (CueO) (Hoegger et al., 2006). Es wurden außerdem Bereiche beschrieben, die durch die Aminosäurereste, die die Kupferionen binden, definiert werden, die stark konserviert sind. Für die Familie der Multikupferoxidasen wurden zwei solcher Sequenzmuster beschrieben (Messerschmidt und Huber, 1990; Solomon et al., 1996), die in ähnlichen Variationen in allen Enzymen dieser Proteinklasse auftreten. Variationen dieser Muster und zwei weitere Sequenzmuster wurden speziell für Laccasen beschrieben. Sie sollten als Unterscheidungsmerkmal von Laccasen und anderen Multikupferoxidasen dienen (Kumar et al., 2003). Diese Sequenzvielfalt und die daraus resultierenden unterschiedlichen biochemischen Eigenschaften der Proteine bei dennoch konserviertem strukturellen Gerüst erlauben die Analysen der Zusammenhänge zwischen Sequenzen, Strukturen und Funktionen der jeweiligen Enzyme. Eine geeigneten Datenstruktur, die die vorhandenen Informationen integriert, fehlte jedoch bislang.

3.3 Systematische Analyse von Proteinfamilien

Techniken zur Sequenzierung kompletter Genome (Venter et al., 2001; Myers et al., 2000) sowie das Forschungsfeld der strukturellen Genomik (Gaasterland, 1998) führten zu einem rasant ansteigenden Volumen an biologischen Sequenz- und Strukturdaten. Die integrierte Sammlung der verfügbaren Daten in einer anwendungsorientierten Datenstruktur, in die die relevanten Analysewerkzeuge implementiert sind, ermöglicht die Analyse von Proteinsequenzen und -strukturen. Als grundlegende Methoden haben sich beispielsweise vergleichende Sequenzanalysen durch Multisequenzalignments (May, 2001), strukturbasierte Sequenzanalysen durch Strukturmultisequenzalignments (Russell und Barton, 1992) und Hidden-Markov-Modelle als probabilistische Vorhersagemethoden (Eddy, 2008) etabliert. Im Kontext einer wohldefinierten Familie lassen sich mit ihnen familienspezifische Deskriptoren ableiten oder konservierte Bereiche auf Sequenz- und Strukturebene identifizieren. Diese können mit Funktionen des Proteins in Verbindung

gebracht werden, wodurch das Verständnis familienspezifischer Eigenschaften unterstützt wird.

3.3.1 Integration biologischer Daten

Die Verfügbarkeit von Daten aus über 165 vollständig sequenzierten Genomen aus zahlreichen Genomprojekten, sowohl aus Eukaryonten als auch aus Prokaryonten, ermöglicht die funktionelle Analyse der Proteine, die durch diese Genomsequenzen codiert werden. Die dadurch entstehenden umfangreiche Datenmengen an biologischen Sequenz- und Strukturinformationen ermöglichen die detaillierte Analyse dieser Proteine. Neue Technologien ermöglichen es, schnell eine große Anzahl an Proteinen zu identifizieren, ihre Funktionen zu bestimmen, Interaktionen festzustellen oder sie in der Zelle zu lokalisieren. Protein-Datenbanken spielen daher eine wichtige Rolle als zentrale Quelle zur Speicherung und Veröffentlichung der Daten (Apweiler et al., 2004). Die Analyse von Proteinsequenzen und -strukturen ist essentiell für die Vorhersage von Funktionen bestimmter Bereiche eines Proteins oder ganzer Proteine. Annotationen, die Proteine oder einzelne Proteinbereiche funktional beschreiben, können jedoch nur erfolgen, wenn ein Verständnis der Sequenz-Struktur-Funktionsbeziehungen gegeben ist. Die Vorhersagen basieren im Allgemeinen auf der Annahme, dass ähnliche Proteine evolutionär miteinander verwandt (homolog) sind, sie deswegen ein gemeinsames Faltungsmuster besitzen und eine ähnliche Funktion haben (Zuckerandl und Pauling, 1965; Chothia und Lesk, 1986). Daher ist es ein vielversprechender Ansatz, Sequenzen und Strukturen von Proteinen im Kontext der homologen Familie, aus der sie hervorgehen, zu untersuchen. Die verlässliche Klassifikation von Proteinen ist essentiell, um eine integrierte Analyse von Sequenz-Struktur-Funktionsbeziehungen durchführen zu können. Die Einteilung kann je nach Proteinklasse aufgrund von Sequenzähnlichkeit und struktureller oder funktionaler Verwandtschaft unterschiedlich erfolgen.

Als Hauptdatenquellen im Internet für Sequenzen haben sich *GenBank* (*Gen Products Data Bank*, (Benson et al., 2008)), *UniProt* (*Universal Protein Resource*, (UniProt Consortium, 2009)) und für Strukturen die *PDB* (*Protein Data Bank*, (Berman et al., 2002)) etabliert. Die umfangreichste hierbei ist die *GenBank*, die insgesamt über 100 Millionen Nukleotid- und Proteinsequenzen enthält (Stand Juli 2009). Trotz der minimalen Annotation bildet die *GenBank* integriert in dem Abfragesystem der *NCBI Entrez* (*National Center of Biotechnology Information*, <http://www.ncbi.nlm.nih.gov>) zusammen mit der biomedizinischen Literaturdatenbank *PubMed* und dem Homologiesuche-Programm

BLAST (*Basic Local Alignment Search Tool* (Altschul et al., 1997)) die wichtigste Daten- und Recherchegrundlage der biomedizinischen Wissenschaft. Die *UniProt/Swissprot* enthält aktuell fast 430 000 Proteinsequenzen (Stand Juli 2009). Darüber hinaus enthält sie Annotationen zu den entsprechenden Proteinen, Querverweise auf andere Datenbanken, Literatureinträge und Funktionszuweisungen. Dagegen beträgt die Anzahl der in der *PDB* hinterlegten Strukturen nur knapp 60 000 (Stand Juli 2009). Diese Lücke erschwert das Verständnis der Proteinfunktionen und des Zusammenhangs zwischen Sequenz-Struktur-Funktionsbeziehung.

3.3.2 Relationale Datenbanken

Das geordnete Speichern von Daten, die meist durch verschiedene Quellen zur Verfügung gestellt werden, in konsistenter nichtredundanter Form wird durch die Integration der Daten in eine Datenbank (DB) realisiert. Damit können spezifisch nur die Daten in Betracht gezogen werden, die für die jeweilige Anwendung von Relevanz sind. Für die effiziente und systematische Analyse von Proteinsequenzen und -strukturen ist eine solche Datenstruktur Grundvoraussetzung. Zur Verwaltung einer Datenbank dient das Datenbankmanagementsystem (DBMS). Das DBMS organisiert die Speicherung und Manipulation der Daten und kontrolliert die Zugriffe auf die Datenbank. Datenbank und Datenbankmanagementsystem bilden zusammen das Datenbanksystem (DBS). Die Datenbankabfragen und die Manipulation der Daten erfolgt über eine Datenbanksprache. Die bekannteste Datenbankabfragesprache ist *SQL* (*Structured Query Language*) (Elmasri und Navathe, 2006).

Die Struktur des Datenbanksystems wird durch das Datenmodell, auf dem es basiert festgelegt. Im relationalen Datenmodell, auf dem die meisten gängigen DBMS basieren, werden die Datenbankobjekte, die modelliert werden sollen, und ihre Beziehungen zueinander durch zweidimensionale Tabellen (Relationen) repräsentiert. Die Spalten der Tabellen beschreiben Attribute der Daten und müssen innerhalb der Tabellen eindeutig benannt werden. Zur eindeutigen Identifizierung eines Datensatzes muss jede Tabelle einen Primärschlüssel (*primary key, PK*) enthalten. Dieser kann aus mehreren Attributen bestehen, muss jedoch eineindeutig definiert sein. Als Fremdschlüssel (*foreign key, FK*) kann er in anderen Tabellen auftreten und dient so der Verknüpfung von Datenbankeinträgen (Codd, 1980).

3.3.3 Data Warehouse System

Als *Data Warehouse* wird ein System bezeichnet, das alle Daten einer untersuchten Domäne in einem lokalen Datenbanksystem integriert und zudem unabhängig von der Verfügbarkeit der öffentlichen Datenquellen ist (Elmasri und Navathe, 2006). Der Vorteil eines solchen Systems besteht darin, dass die heterogenen Daten der Quelldatenbanken durch die Integration in ein einheitliches, konsistentes Datenmodell und in ein ebenfalls einheitliches Format gebracht werden. Die Entwicklung eines solchen Datenmodells ist der essentielle Schritt bei der Erstellung eines *Data Warehouse*. Weiterhin müssen die Daten aus den Quelldatenbanken extrahiert und in ein einheitliches Format transformiert werden. Diese Schritte sind innerhalb eines *Data Warehouse* in der so genannten *ETL layer* (*extraction, transformation, load*) implementiert (Abbildung 3.3.1).

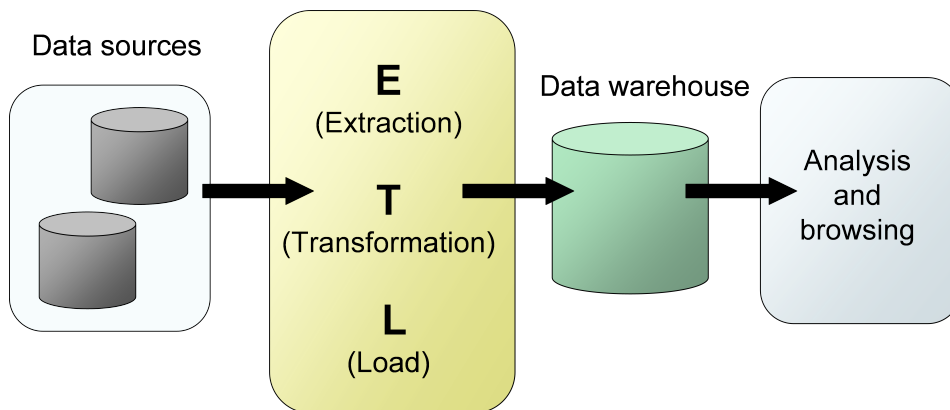


Abbildung 3.3.1: Zentrale Komponente des *Data Warehouse*: Die *ETL-Layer*.

Am Institut für Technische Biochemie der Universität Stuttgart wurde eigens für die Integration von Proteinfamilien und deren systematische Analyse das Datenbanksystem *DWARF* (*Data Warehouse for Analyzing Protein Families*, (Fischer et al., 2006)) entwickelt. Dieses erlaubt es, Sequenz- und Strukturdaten in einem lokalen, relationalen Datenbanksystem in nichtredundanter, konsistenter Form abzulegen, und sie dabei in Superfamilien und homologe Familien zu klassifizieren. Das *DWARF*-System stellt zudem zahlreiche Analysewerkzeuge zur Verfügung, die beliebig erweitert werden können. Innerhalb dieses Datenbanksystems wurden bereits mehrere Datenbanken erstellt (Fischer und Pleiss, 2003) und erfolgreich analysiert (Knoll und Pleiss, 2008; Knoll et al., 2009; Widmann et al., 2008).

3.3.4 Vergleichende Sequenzanalyse

Die häufigste Methode in der Bioinformatik und der molekularen Phylogenie zur Analyse funktioneller oder evolutionärer Verwandtschaften, genannt Homologie, von Nukleotid- oder Proteinsequenzen ist ein Sequenzalignment. Dabei werden die Sequenzen positionsweise verglichen und identische oder möglichst ähnliche Elemente einander zugeordnet unter der Annahme, dass viele gleiche oder ähnliche Elemente in gleicher Reihenfolge auf eine evolutionäre oder funktionelle Verwandtschaft hinweisen. Eine Fehlpaarung in einem solchen Alignment kann auf eine Mutation hindeuten. Um ein möglichst sinnvolles Alignment zu erhalten, dürfen *gaps* (Leerstellen, Lücken) eingefügt oder erweitert werden. Evolutionär bedeuten *gaps* eine Deletion oder Insertion. Die Ähnlichkeit der Elemente wird durch eine sogenannte *scoring*-Matrix bestimmt. In diese gehen die Aminosäureeigenschaften und Mutationswahrscheinlichkeiten ein (Durbin et al., 1998). Ein Beispiel für das Alignment zweier Zeichenketten ist:

```
IMISSMISSISSIPPI      I-MISSMISSISSIPPI-
                        | | | | | | | | | | | | | | | | | | | | | | | |
MYMISSISAHIPPIE       MYMISS-ISAH--IPPIE
```

Um die durch die Evolution am wahrscheinlichsten entstandene Paarung zu erhalten, berücksichtigen Alignment-Algorithmen eine Kostenfunktion, die durch ein optimales Alignment minimiert wird. Die ersten Arbeiten zur Informationsgewinnung über evolutionäre Zusammenhänge homologer Proteine durch den direkten Vergleich ihrer Aminosäurezusammensetzung gehen bis in die 60er Jahre zurück (Zuckerandl und Pauling, 1965). Der von Needleman und Wunsch (1970) entwickelte Algorithmus, der durch ein dynamisches Programm das optimale paarweise Alignment entwickelt, war über 20 Jahre lang das Arbeitspferd für die vergleichende Analyse von Nukleotid- und Proteinsequenzen (Doolittle, 1994). Er wird auch heute noch als Standardmethode angewandt, um den Grad der Verwandtschaft, und damit die Ähnlichkeit zwischen zwei Sequenzen, und damit die Homologie, zu ermitteln und ist zudem die Grundlage für viele Alignment-Programme.

Das optimale paarweise Alignment kann exakt in polynomieller Laufzeit $\mathcal{O}(nm)$, wobei n und m die Längen der zu vergleichenden Sequenzen sind, berechnet werden. Die Laufzeit steigt beim multiplen Sequenzalignment (MSA), also beim Vergleich mehrerer Sequenzen, exponentiell mit der Anzahl der zu vergleichenden Sequenzen. Um jedoch aus Proteinfamilien ein biologisch bzw. evolutionär sinnvolles Alignment zu berechnen, sowie konservierte Positionen zu analysieren und familienspezifische Parameter daraus zu

gewinnen, werden Alignments aller homologen Mitglieder einer Proteinfamilie benötigt. Um den Rechenaufwand einzuschränken, wurden Heuristiken entwickelt, die bei deutlich geringerem Rechenaufwand eine Lösung liefern, welche möglichst nahe an der optimalen Lösung liegt. Der Feng-Doolittle-Algorithmus verwendet eine progressive Strategie, bei der hierarchisch und basierend auf einer Clusteranalyse schrittweise ein multiples Alignment ermittelt wird (Feng und Doolittle, 1987). Dieser ist heute in erweiterter Form in dem Standard-Programm für multiple Sequenzvergleiche *CLUSTAL W* implementiert (Thompson et al., 1994).

Um Alignments entfernter verwandter Proteinsequenzen zu verbessern, werden häufig hybride Methoden eingesetzt, die zusätzlich noch strukturelle Eigenschaften mit einbeziehen. Diese beruhen auf der Annahme, dass selbst entfernt verwandte Proteine sehr ähnliche 3D-Strukturen besitzen (Gotoh, 1996). Verbreitete Implementierungen solcher strukturbasierter Alignment-Methoden sind beispielsweise *3DCoffee* (O’Sullivan et al., 2004) oder *STAMP* (*Structural Alignment of Multiple Proteins*, Russell und Barton (1992)).

3.3.5 Probabilistische Modellierung

Das Konservierungsmuster verwandter Proteinsequenzen kann durch multiple Sequenzalignments analysiert werden. In individuellen Fällen ist eine positionsspezifische Information erwünscht. Um aus einem multiplen Sequenzalignment durch eine statistische Modellierung positionsspezifische Informationen zu erhalten, werden Profil-*HMMs* (*Hidden Markov* Modell, (Eddy, 1998)) eingesetzt. Ein umfangreiches Programmpaket für *HMM*-Profile bildet *HMMER* (Eddy, 1998). Wie auch alle anderen *HMM*-Profil-Modelle basiert es auf der *HMM*-Theorie (Durbin et al., 1998). *HMMs* sind probabilistische Modelle, die auf Markovketten beruhen. Eine Markovkette ist ein stochastischer Prozess, bei dem zu jedem Zeitpunkt die Wahrscheinlichkeiten aller zukünftigen Zustände vom momentanen Zustand abhängen. Die Markovkette wird daher durch Zustände und Übergangswahrscheinlichkeiten beschrieben. Die Zustände der Kette sind nicht sichtbar (*hidden*) und es werden gemäß einer zustandsabhängigen Wahrscheinlichkeitsverteilung nach außen sichtbare Symbole angezeigt.

HMM-Profile bilden eine wichtige Grundlage für Homologiesuchen entfernt verwandter Proteine (Karplus et al., 1998). Aus einem strukturbasierten Alignment generiert können sie für Proteinfaltungsvorhersagen (*fold recognition*) verwendet werden und somit struk-

turelle Informationen über funktionell relevante Positionen in Proteinen liefern (Scheeff und Bourne, 2006). Dadurch kann selbst für Proteine ohne bekannte Kristallstruktur ein tieferes Verständnis in Sequenz-Funktionszusammenhänge erlangt werden.

3.3.6 Von der Sequenz zur Funktion

Durch die Etablierung einer familienspezifischen Datenbank können große Mengen an Sequenz- und Strukturinformationen organisiert abgelegt werden. Eine solche Datenbank, die außerdem die bisher beschriebenen Analysewerkzeuge integriert, dient als Hilfsmittel, um systematische Analysen durch umfassende Sequenz- und Strukturvergleiche durchzuführen und daraus familienspezifische Eigenschaften abzuleiten. Dies wurde bereits im Zusammenhang verschiedener Proteinfamilien, wie zum Beispiel für die Proteinklasse der Lipasen, der PHA-Depolymerasen und der mittelkettigen Dehydrogenasen/Reduktasen (Fischer und Pleiss, 2003; Knoll et al., 2009; Knoll und Pleiss, 2008) erfolgreich gezeigt. Die Klassifikation nach Sequenzähnlichkeit und basierend auf phylogenetischen Analysen erlaubt die Erstellung verlässlicher Multisequenzalignments, aus denen wiederum familienspezifische *HMM*-Profile abgeleitet werden können. Durch Multisequenzalignments konnten funktionell relevante Aminosäuren identifiziert und annotiert werden. Durch die gezielte Mutation solcher Positionen können Enzyme mit optimierten Eigenschaften hergestellt werden. So konnten durch den datenbankgestützten systematischen Sequenz- und Strukturvergleich von Cytochrom P450-Monooxygenasen so genannte *hot-spot*-Positionen identifiziert werden (Seifert und Pleiss, 2008), auf deren Basis eine Mutantenbibliothek für eine bakterielle Cytochrom P450-Monooxygenase erstellt wurde, die sich durch erhöhte Selektivität auszeichnet (Seifert et al., 2009).

4 Ziele dieser Arbeit

Durch die systematische Analysen von Sequenzen und Strukturen von Proteinfamilien, die auf Grund ihrer katalytischen Eigenschaften für biotechnologische Applikationen sehr interessant sind, sollten familienspezifische Funktionsparameter abgeleitet werden. Zu diesen Proteinen zählen die Familien der Cytochrom P450-Monooxygenasen und der Multikupferoxidasen, insbesondere die Unterfamilie der Laccasen. Der essentielle Schritt für sequenz- und strukturumfassende Analysen zur Ableitung von familienspezifischen Proteineigenschaften ist die Integration der relevanten Daten einer Proteinfamilie in einem konsistent organisierten, nicht-redundanten Datensystem. Daher sollten Proteinfamilien-Datenbanken erstellt werden, die Sequenz-, Struktur- und funktionelle Informationen mit entsprechenden Analysewerkzeugen integrieren. Die Erkenntnisse der Analysen sollten Rückschlüsse auf die Sequenz-Struktur-Funktionsbeziehungen geben, auf deren Basis gezielt Mutationen vorhergesagt werden können, die eine optimierte Anwendung der Enzyme ermöglichen.

Für die Familie der Cytochrom P450-Monooxygenasen sollte eine familienspezifische Datenbank erstellt werden, in der alle verfügbaren Sequenz- und Strukturinformationen strukturiert abgelegt werden können. Diese Datenbank sollte systematische Analysen von Sequenz und Struktur möglich machen, auf deren Basis das Verständnis der Struktur-Funktion-Sequenz-Zusammenhänge erweitert werden sollte. Die ständig neu hinzukommenden Datenmengen und die für verschiedene Problemstellungen unterschiedlichen Anforderungen an eine solche Datenbank erforderten außerdem die kontinuierliche Aktualisierung der Daten und die Erweiterung der Funktionalitäten, wie zum Beispiel die Integration biochemischer Eigenschaften und die Vorhersage und Annotation von strukturell konservierten Bereichen in Sequenzen von Cytochrom P450-Monooxygenasen (S. 60, S. 63, S. 71). Durch die strukturelle Vorhersage sollte die Analyse funktionell relevanter Positionen und Regionen, die sich in hochvariablen Sequenzregionen befinden, ermöglicht werden, um daraus biochemische Eigenschaften abzuleiten.

Mit Hilfe bioinformatischer Modellierungen sollten experimentelle Phänomene bei der Immobilisierung von Cytochrom P450 BM-3 auf mesoporösen Materialien erklärt werden. Durch diese Modelle sollte die Abhängigkeit der Immobilisierungseffizienz von Eigenschaften wie Größe des Proteins, dessen elektrostatischen Potentials und die pH-Abhängigkeit von Oberflächenladungen erklärt werden. Außerdem sollte darauf basierend die Orientierung des Proteins auf der Oberfläche vorhergesagt werden (S. 97).

Um ein besseres Verständnis der Sequenz-Struktur-Funktionsbeziehungen der Laccasen innerhalb der Familie der Multikupferoxidasen zu ermöglichen, sollte eine umfassende Datenbank für die gesamte Proteinfamilie erstellt werden, die eine systematische Einteilung der Proteine beinhaltet. Die gesamte Proteinfamilie sollte im Hinblick auf Sequenzähnlichkeit, Organismen und Funktion untersucht werden. Die für die Funktion entscheidenden Sequenzbereiche sollten erfasst, annotiert und analysiert werden, um ein Werkzeug für systematische Analysen familienspezifischer Eigenschaften der gesamten Proteinfamilie zu generieren (S. 106).

5 Ergebnisse und Diskussion

5.1 Cytochrom P450-Monooxygenasen

5.1.1 Die Cytochrome P450 Engineering Database: Ein Orientierungs- und Vorhersagewerkzeug für die Cytochrom P450 Proteinfamilie

(siehe: *The Cytochrome P450 Engineering Database: A Navigation and Prediction Tool for the Cytochrome P450 Protein Family*, Seite 60)

Die große und diverse Familie der Cytochrom P450-Monooxygenasen (CYPs) umfasst Häm-Proteine, die in fast allen Organismen vertreten sind und eine Vielzahl chemischer Reaktionen katalysieren (Montellano, 1995). Beim Menschen sind sie hauptsächlich an der Umsetzung von Medikamenten oder Schadstoffen beteiligt. Daher ist ein Verständnis der Zusammenhänge zwischen Sequenz-Struktur-Funktion der Proteine dieser Familie von besonders großer Bedeutung (Raucy und Allen, 2001). Obwohl es schon zahlreiche Quellen gibt, die *online* Informationen über CYPs zur Verfügung stellen, fehlt es an einer einheitlich strukturierten Datenbank, die die relevanten Sequenz-, Struktur- und Funktionsdaten integriert. Daher wurde die *Cytochrome P450 Engineering Database (CYPED)* implementiert, um systematische und umfassende Vergleiche der Proteinsequenzen und Strukturen innerhalb der diversen Familie der Cytochrom P450-Monooxygenasen zu ermöglichen.

Als Grundlage für die Erstellung der Datenbank wurde das *Data Warehouse System DWARF* (Fischer et al., 2006) verwendet. Das *DWARF*-System basiert auf einem relationalen Datenmodell, das für die Analyse von Sequenz-Struktur-Funktions-Beziehungen von Proteinfamilien erstellt wurde. Mit Startsequenzen von der Cytochrome P450 Homepage (Nelson, 2002) wurden per *BLAST*-Suche (Altschul et al., 1997) homologe Sequenzen, Strukturen und Annotationsinformationen aus der *GenBank* (Benson et al., 2008) und

der *PDB* (Berman et al., 2002) extrahiert und in dem Datenbanksystem gespeichert. Alle Proteine wurden in homologe Familien und in Superfamilien gemäß der Standardklassifizierung nach Nelson (Nelson, 2006) eingeteilt. Für Sequenzen, zu denen Struktureinträge vorhanden waren, wurden Sekundärstrukturen mittels *DSSP* (Kabsch und Sander, 1983) berechnet und in der Datenbank annotiert. Weiterhin wurden funktionell relevante Aminosäuren und Motive aus den *GenBank*-Einträgen und aus der Literatur (Kemper, 2004; Mestres, 2005) innerhalb der Multisequenz-Alignments annotiert. Diese Annotationen wurden automatisch auf konservierte Aminosäuren innerhalb der Familien übertragen und lieferten damit Informationen für wenig charakterisierte CYPs.

Die erstmals auf <http://www.cyped.uni-stuttgart.de> veröffentlichte Version der *CYPED* enthält Sequenzinformationen von 3911 Proteinen. Für 25 verschiedene Proteine aus 20 homologen Familien sind Kristallstrukturen abgelegt. Die Proteineinträge sind unterteilt in 1111 homologe Familien und 531 Superfamilien, die als *CLUSTAL W*-Multisequenzalignments (Thompson et al., 1994) repräsentiert werden können. Der Zugriff erfolgt auf Ebene der Sequenzfamilien, Organismen oder Strukturen. Außerdem ist für jede Familie der zugehörige phylogenetische Baum und ein *HMM*-Profil abgelegt. Die jeweiligen Einträge sind mit den Quelleinträgen der *GenBank* verlinkt. Über eine Plattform für *BLAST*-Suchen (Altschul et al., 1997) innerhalb der *CYPED* können neue oder unbekannte Sequenzen klassifiziert werden. Der gesamte Datenbankinhalt kann über die Seite heruntergeladen und in eigene Anwendungen integriert werden. Die automatische *update*-Funktion sorgt für die stetige Aktualität der Daten, für die Integration und die Klassifizierung neuer Sequenzen und erweiterter oder geänderter Annotationsinformationen.

Damit bildet die *CYPED* im Gegensatz zu den bisherigen Cytochrom P450 Datenquellen die erste Datenbank, die durch die lokale Integration von Sequenz, Struktur und funktionellen Annotationen und den entsprechenden Methoden systematische Analysen von Sequenz und Struktur der Cytochrom P450-Monooxygenasen erlaubt. Das durch eine solche Analyse neu erlangte Wissen unterstützt das Gesamtverständnis der Funktionsweise dieser Proteine und ermöglicht das Design von Proteinen mit verbesserten biochemischen Eigenschaften, wie in den folgenden Kapiteln erläutert wird.

5.1.2 Die Cytochrome P450 Engineering Database: Integration biochemischer Eigenschaften

(siehe: *The Cytochrome P450 Engineering Database: Integration of Biochemical Properties*, Seite 63)

Nach der Veröffentlichung *CYPED* (Fischer et al., 2007) als Analysewerkzeug für umfassende und systematische Analysen der Cytochrom P450-Monooxygenasen, wurde sie erfolgreich angewandt, um selektivitäts- und spezifizitätsbestimmende Aminosäuren zu bestimmen (Seifert und Pleiss, 2008), sowie die Familien in der *Fungal Cytochrome P450 Database* anzupassen (Park et al., 2008) und diente als Vorlage für Datenbanken anderer Proteinfamilien (Knoll et al., 2009; Knoll und Pleiss, 2008). Die Anzahl der Proteine hat sich seither mehr als verdoppelt, außerdem sind inzwischen Kristallstrukturen zu fast 40 Proteinen bekannt. Daher wurde die im erweiterbaren *DWARF*-System (Fischer et al., 2006) bestehende Datenbank neben der Integration neuer Sequenz- und Strukturinformationen um Daten über biochemische Eigenschaften und neue Funktionalitäten ergänzt.

Die Implementierung dieser Version beinhaltet einen im Gegensatz zur vorherigen Version leicht modifizierten Algorithmus zur Sequenzauswahl. Bei der ersten Version wurden zunächst die Datenbank mit Sequenzen befüllt wurde und diese dann basierend auf Sequenzhomologien und der Einteilung nach Nelson in Familien unterteilt. Bei der neuen Version wurden zuerst die Familien entsprechend Nelsons *Cytochrome P450 Homepage* (Nelson, 2002) erstellt. Aus der letzten Version der *CYPED* wurden die konsistent benannten Einträge als Startsequenzen verwendet, um die erstellten Familien zu befüllen. Dadurch konnte basierend auf fast 400 Startsequenzen ein Ergebnis von 8613 Proteinen, zu denen 47 Strukturen aus 36 homologen Familien bekannt sind, erzielt werden. Diese Proteine sind auf 619 homologe Familien und 249 Superfamilien verteilt. Die geringe Familienanzahl bei mehr als doppelt so vielen Proteinen wie in der ursprünglichen Version der *CYPED* beruht auf dem systematischeren Ansatz, durch den Familien mit sehr wenigen Mitgliedern vermieden wurden.

Die neue Version der *CYPED* zeigt jeden Proteineintrag über eine *feature page* an. Auf dieser werden die Sequenz und ihre Annotationen dargestellt (Abbildung 5.1.1). Eine neu entwickelte dynamische Webschnittstelle zeigt Änderungen in der Datenbank zeitgleich an. Weiterhin beinhaltet die *CYPED* neue Funktionen:

- Von der *Home Page of Human Cytochrome P450 (CYP) Allele Nomenclature Committee* (Oscarson und Ingelman-Sundberg, 2002) wurden Informationen über humane CYP Allele extrahiert und in eigens dafür erstellten Tabellen abgespeichert. Die Mutationen und ihre Effekte, also ob das Enzym an Aktivität verloren oder eine erhöhte Aktivität gewonnen hat, sind auf der *feature page* aufgelistet.
- Für alle Einträge der *CYPED* wurden durch ein neu entwickeltes Verfahren, das ein strukturbasiertes *HMM*-Profil in einem automatischen Annotationsprogramm integriert, die konservierten Sekundärstrukturen vorhergesagt. Sie wurden als Annotationen im *DWARF*-System gespeichert und auf den Proteinsequenzen der *feature page* und der Alignments aufgeblendet.
- In Zusammenarbeit mit dem *Institute of Biomedical Chemistry Moscow* konnten die Datenbankeinträge mit denen der *Cytochrome P450 Knowledgebase (CPK)*, (Lisitsa et al., 2001)) verknüpft werden, sofern sie in dieser vorhanden sind. Zu diesem Zweck wurde von Florian Wagner ein neuer Algorithmus implementiert, der auf Sequenzähnlichkeiten als metrischer Primärschlüssel basiert. Diese Definition war erforderlich, da die Integration von Proteindatenbanken auf der Grundlage eines gemeinsamen eindeutigen Schlüssels erfolgt. Bei einem Proteineintrag kann lediglich die Proteinsequenz selbst als datenbankunabhängiges Attribut, das einem Primärschlüssel entspricht, verwendet werden. Ein Primärschlüssel muss allerdings spezifisch für einen Eintrag sein, während die Proteinsequenz von ein und demselben Proteineintrag, beispielsweise durch Mutationen, leicht variieren kann. Daher wurde in einem mehrstufigen Verfahren über eine *BLAST*-Suche (Altschul et al., 1997) in der *CPK* und einem paarweisen globalen Alignment (Needleman und Wunsch, 1970) zu jedem Eintrag der *CYPED*, die entsprechende Verknüpfung erstellt. Neben den Annotationen der für Struktur und Funktion relevanten Regionen, enthält nun jeder *CYPED* Eintrag über die Schnittstelle zur *CPK* zusätzliche Informationen über biochemische Eigenschaften, Substrate, Inhibitoren und Induktoren.

Die *CYPED* stellt eine Reihe nützlicher Methoden, wie zum Beispiel familienspezifische annotierte Multisequenzalignments, phylogenetische Bäume, *HMM*-Profile und eine Schnittstelle für die *BLAST*-Abfrage zur Analyse und Klassifikation der Proteine der Familie der Cytochrom P450-Monooxygenasen zur Verfügung. Um tiefere Einblicke in die Funktionsweise dieser Proteine zu gewinnen, wurde die *CYPED* um neue Funktionalitäten

erweitert. Die biochemischen Informationen, die über die *CPK*-Verlinkung integriert wurden, ermöglichen ein erweitertes Verständnis der Sequenz-Struktur-Funktionsbeziehungen.

The screenshot displays the feature page for [CYP1A1] cytochrome P450 monooxygenase CYP1A1. It includes a reference sequence with amino acid positions 80, 160, 240, 320, 400, 480, and 560. A table titled 'Link to Cytochrome P450 Knowledgebase (CPK)' lists three entries: 1A0_SPARUS_AURATA (92.71% identity), 1A0_STENOTOMUS_CHRYSOPS (91.75% identity), and 1A0_LIMANDA_LIMANDA (84.45% identity). A pop-up window titled 'Functional properties for [1A0~SPARUS AURATA]' shows 'INDUCERS (1 entries)' with 2,3,7,8-TETRACHLORODIBENZO-P-DIOXIN (TCDD; 2,3,7,8-TCDD; DIOXIN) listed as an inducer.

Abbildung 5.1.1: Die *feature page* der *CYPED* zeigt für jedes Protein in der Datenbank die Sequenz, ihre vorhergesagte Sekundärstruktur und die entsprechende Verknüpfung zur *CPK* an.

Daher wurde mit der neuen Version der *CYPED* ein noch leistungsfähigeres System generiert, das die Navigation im CYP-Sequenzraum, in ihren Strukturen und die Analyse und Vorhersage der Proteinfunktionen erlaubt. Sie wurde bei der Bestimmung selektivitäts- und spezifizitätsbestimmender Aminosäuren und dem daraus resultierenden Design von Mutanten (Seifert und Pleiss, 2008; Seifert et al., 2009) verwendet. Außerdem wurde im Rahmen dieser Arbeit basierend auf den Daten der *CYPED* ein Vorhersagewerkzeug für die konservierten Sekundärstrukturen entwickelt. Die Implementierung und die daraus resultierende Analyse der für CYP102A1 aus *Bacillus megaterium* (P450 BM-3) aktivitäts- und selektivitätsbestimmenden Position 87, sowie die Analyse der Sequenzbereiche, die die Reduktaseinteraktionsfläche ausmachen, wird im folgenden Kapitel beschrieben.

5.1.3 Die modulare Struktur von Cytochrom P450-Monooxygenasen

(siehe: *The Modular Structure of Cytochrome P450 Monooxygenases*, Seite 71)

Während CYP-Sequenzen innerhalb einer Superfamilie eine Sequenzidentität von mindestens 40 % aufweisen, beträgt die Sequenzidentität von Sequenzen unterschiedlicher Superfamilien häufig lediglich 15-20 % (Graham und Peterson, 1999). Trotz dieser be-

trächtlichen Sequenzdiversität sind die Strukturen der CYPs in ihren Sekundärstrukturen und ihrer Faltung hochkonserviert. Schon frühere Analysen haben gezeigt: CYPs bestehen aus konservierten Regionen, die für Struktur und Funktion essentiell sind und aus variablen Regionen, die die individuellen biochemischen Eigenschaften bestimmen (Mestres, 2005). Die definierten Sekundärstrukturen werden als Helices αA bis αK und als Faltblätter $\beta 1$ bis $\beta 4$ bezeichnet, die zusammen die so genannte CYP-Faltung darstellen (Peterson und Graham, 1998). Außer den strukturell konservierten Bereichen wurden Regionen beschrieben, die sowohl in Sequenz als auch in ihrer Struktur variieren. Sechs Bereiche, die an der Substraterkennung und -bindung beteiligt sind, wurden als SRS1 - SRS6 definiert (Gotoh, 1992). In der SRS1, die sich im hochvariablen BC-loop befindet, ist ein Aminosäurerest lokalisiert, der direkt zum Häm zeigt und in vorigen Arbeiten als aktivitäts- und selektivitätsbestimmend beschrieben wurde. In CYP102A1 aus *Bacillus megaterium* (P450 BM-3) ist dieser Rest das an der Position 87 lokalisierte Phenylalanin. Zahlreiche Mutationen von F87 mit Einfluß auf Regio- und Stereoselektivität bei verschiedenen Substraten wurden bei P450 BM-3 bereits durchgeführt (Li et al., 2008; Urlacher und Schmid, 2002; Urlacher et al., 2006; Seifert und Pleiss, 2008). Die Identifikation der konservierten Strukturbereiche (SCR, *structurally conserved region*) und von Aminosäureresten mit Schlüsselfunktionen ist von großem Vorteil beim Design von Proteinen mit verbesserten Eigenschaften.

Die gemeinsamen strukturellen Eigenschaften von CYPs machten es möglich, aus 31 Proteinstrukturen ein Modell abzuleiten, mit dem die strukturell konservierten Regionen (SCR) der CYPs auf Proteine übertragen werden konnten, für die noch keine Strukturinformation vorliegt. Dazu wurde zunächst ein Strukturalignment erstellt, aus dem ein verlässliches Struktur-HMM-Profil generiert werden konnte. Mit Hilfe des *DWARF*-Systems wurden SCR in den Multisequenzalignments und auf den *feature page* aller 8614 Proteineinträge der *CYPED* annotiert und zusammen mit der *CYPED* (Sirim et al., 2009) veröffentlicht. Die Vorhersage der SCRs kann auf <http://www.cyped.uni-stuttgart.de/cgi-bin/strpred/dosecpred.pl> für jede beliebige CYP-Sequenz durchgeführt werden. Damit können SCRs für neu identifizierte Sequenzen vorhergesagt werden, auch wenn sie noch nicht in der *CYPED* integriert sind.

Eine strukturelle Analyse der BC-loop Region (SRS1) zeigte die deutlichen Unterschiede der loops verschiedener Kristallstrukturen. Trotzdem konnte in fast jeder Struktur ein Aminosäurerest identifiziert werden, der analog zu P450 BM-3 F87 direkt zum Häm zeigt.

Mit Hilfe des Struktur-*HMM*-Profils konnte diese Position in 80 % aller analysierten Strukturen korrekt vorausgesagt werden, in den übrigen Strukturen mit einer maximalen Abweichung von 2 Positionen. Durch die Anwendung auf 11 neue CYP-Strukturen konnte die Genauigkeit der Methode validiert und bestätigt werden. Dies erlaubte eine analoge Analyse dieser Position in allen 8614 Proteine der *CYPED*. Es konnte gezeigt werden, dass 73 % der Aminosäurereste, die für die Position vorhergesagt wurden, von aliphatischer Natur oder ein Phenylalanin sind. Weitere 24 % bestehen aus kleinen, polaren Aminosäuren und nur insgesamt 3 % aus geladenen Aminosäuren.

Die zur Aktivierung des Luftsauerstoffs benötigten zwei Elektronen werden CYPs in den meisten Fällen von einem Redox-Partner übertragen. CYPs können daher auch nach Elektronentransfersystemen eingeteilt werden (McLean et al., 2005). Eine Untersuchung der Reduktase-Interaktionsregionen (Hasemann et al., 1995) zeigte Unterschiede in Länge und Konformation der jeweiligen Strukturbereiche. Die Identifikation zweier solcher *reductase interaction sites* (RIS1 und RIS2) in allen *CYPED* Sequenzen erlaubte die Analyse der Häufigkeit der unterschiedlichen Längen. Dabei konnten die Proteine, die sich im Hinblick ihrer Redox-Partner unterscheiden, deutlich voneinander durch ihre Längen getrennt werden. Den größten Anteil bildeten die Proteine mit langer RIS1 und RIS2 und vorwiegend zu der CYP-Klasse I gehören, die mit einer FMN/FAD Cytochrome P450-Reduktase (CPR) interagieren. Weniger häufig traten Proteine mit kurzen RIS auf. Sie werden der Klasse II zugeordnet, die mit anderen Redox-Partnern interagieren. Den kleinsten Prozentsatz bildeten Proteine mit langer RIS1 und sehr langer RIS2. Unter ihnen wurden Proteine identifiziert, die keinen Redox-Partner benötigen. Die Interaktion von CYPs mit ihren Redox-Partnern ist eine wichtige Voraussetzung für das Stattfinden der Reaktion, jedoch bisher weitgehend ungeklärt. Das Verständnis der Faktoren, die die Interaktion von CYPs mit potentiellen Redox-Partnern beeinflussen wurde durch diese Analyse verbessert und kann somit das Design von Systemen mit optimierten Interaktionen unterstützen.

Die hier beschriebene Methodik erlaubt eine strukturelle Navigation in CYP Sequenzen, auch wenn keine Strukturinformationen vorliegen. Außerdem konnten Aminosäurereste an funktionellen Schlüsselpositionen vorhergesagt werden. Damit wird ein sehr hilfreiches Werkzeug zur Verfügung gestellt, mit dem für *protein engineering* vielversprechende Angriffspunkte, wie zum Beispiel die selektivitäts- und aktivitätsbestimmende Position, die untersucht wurde, vorhergesagt werden können.

5.1.4 Modellierung von P450 BM-3 für Immobilisierungsversuche

(siehe: *Immobilization of P450 BM-3 on mesoporous molecular sieves*, Seite 97)

Die Immobilisierung von Cytochrom P450-Monooxygenasen auf verschiedenen Trägermaterialien kann eine deutliche Steigerung der Aktivität und der Stabilität der Proteine bewirken. Mesoporöse Materialien sind auch bei größeren Proteinen anwendbar. Daher wurden Proteineigenschaften bestimmt, die für die Interaktion mit dem Trägermaterial entscheidend sind, wie pH-Abhängigkeit, elektrostatisches Potential zur Bestimmung der Orientierung und einfache Struktureigenschaften wie Molekülgröße, um einen Anhaltspunkt für die Porengröße des zu verwendenden Materials zu bestimmen (Hudson et al., 2005; Trodler et al., 2008; Yoon und Lenhoff, 1992).

Die theoretischen Berechnungen wurden analog der Experimente, bei denen die Häm-Domäne von Cytochrom P450 aus *Bacillus megaterium* (P450 BM-3) auf mesoporösen Aluminium-Silikaten immobilisiert wurde, durchgeführt. Die experimentiellen Immobilisierungsarbeiten wurden innerhalb der Arbeitsgruppe Urlacher am Institut für Technische Biochemie der Universität Stuttgart durchgeführt. Die Synthese der Materialien erfolgte innerhalb der Arbeitsgruppen Hunger, Institut für Technische Chemie, Universität Stuttgart und Gläser, Institut für Technische Chemie, Universität Leipzig.

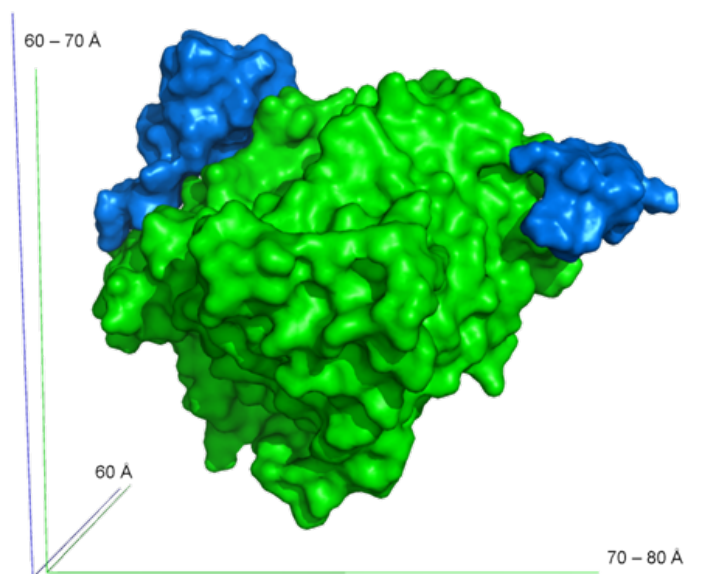


Abbildung 5.1.2: Molekulare Dimensionen der Häm-Domäne von P450 BM-3. Die Regionen der Kristallstruktur (PDB: 1BU7) sind in grün dargestellt, die modellierten Teile in blau. Die Größe des modellierten Proteins beträgt $80 \times 70 \times 60 \text{ \AA}^3$.

Um eine vollständige Struktur, die die noch nicht kristallographisch erschlossene *linker*-Region und das in den Experimenten verwendete *His6-Tag* enthält, zu bestimmen, wurde ein Homologiemodell (Sali und Blundell, 1993) erstellt. Dazu wurden die PDB-Strukturen der Häm-Domäne von P450 BM-3 und der Häm-Domäne fusioniert mit der FMN-Domäne als Templat verwendet. Die weitere molekulardynamische Simulation ergab eine realistischere Konformation des Modells (Walker et al., 2008; Giammona et al., 1984; Wang et al., 2000; Cramer und Truhlar, 1999). Die mit Hilfe von *VMD* (Humphrey et al., 1996) bestimmte Größe des Proteins beträgt $80 \times 70 \times 60 \text{ \AA}$, also etwas größer als die ursprünglichen $70 \times 60 \times 60 \text{ \AA}$ ohne den *linker* und dem *His6-Tag* (Abbildung 5.1.2). Weiterhin wurde eine Titrationskurve für die pH-Werte zwischen 3 und 10 berechnet (Alexov und Gunner, 1997). Daraus ergab sich, dass die Gesamtladung des Proteins von 32 bis -18, bei ansteigendem pH-Wert abnimmt. Der damit bestimmte *pI* liegt bei 5,4. Bei einem pH-Wert von 7 ist das Protein negativ geladen und hat eine Gesamtladung von -7,9 (Abbildung 5.1.3). Bei diesem pH-Wert konnte die größte Menge an Protein immobilisiert werden.

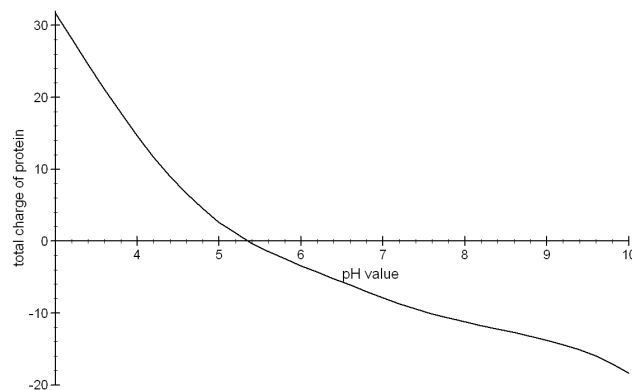


Abbildung 5.1.3: Die berechnete Titrationskurve des erweiterten Strukturmodells von P450 BM-3 zeigt eine bei ansteigendem pH Wert zunehmend negativ werdende Gesamtladung des Proteins mit einem *pI* von 5,4.

Die Berechnungen des elektrostatischen Potentials für die pH-Werte 6, 7 und 8 (Nicholls und Honig, 1991; Sitkoff et al., 1994) zeigten, dass das Protein bei allen drei pH-Werten negativ geladen ist und einen positiven *patch* besitzt. Die Fläche dieses *patch* ist bei pH 6 am größten und nimmt mit steigendem pH-Wert ab. Er liegt in der proximalen Region, die höchstwahrscheinlich mit der Reduktase-Domäne interagiert (Sevrioukova et al., 1999). Die gegenüberliegende Seite (distal) mit dem Eintrittskanal für das Substrat in das aktive Zentrum (Li und Poulos, 1997), ist negativ geladen (Abbildung 5.1.4).

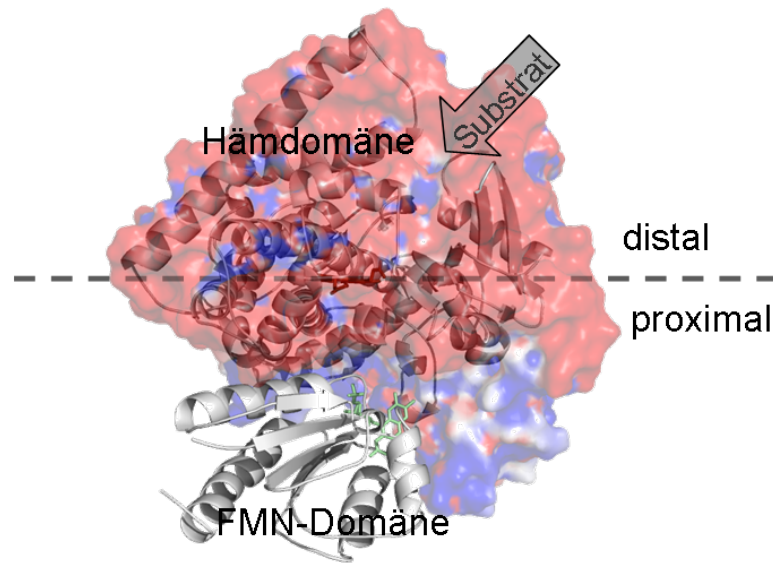


Abbildung 5.1.4: Elektrostatische Ladungsverteilung bei pH 7 an der Oberfläche des Strukturmodells der dreidimensionalen Struktur der Häm-Domäne und der FMN-Domäne von P450 BM-3 aus *Bacillus megaterium* (basierend auf den PDB-Einträgen 1BVY und 1BU7), in *surface*-Darstellung. Die Ladungsverteilung zeigt in proximaler Region einen positiven *patch* auf der Häm-Domäne, während die Oberfläche an der distalen Seite, wo sich auch der Substrateingangskanal befindet, negativ geladen ist. Häm- und FMN-Domäne des Proteins sind in *cartoon* dargestellt.

Durch die Modellierung der biochemischen Eigenschaften der Häm-Domäne von Cytochrom P450 BM-3 war eine molekulare Interpretation der optimalen Immobilisierungsbedingungen möglich, die von pH-Wert, Porengröße des Materials etc. abhängen. Dies ließ darauf schließen, dass in Hinblick auf Protein- und Porengröße eine Bindung des Proteins innerhalb der Poren möglich sein müsste (Hernández et al., 2005). Dennoch deuten die experimentellen Ergebnisse darauf hin, dass die Diffusion in die Poren behindert wird. Dies führt zu der Annahme, dass die Porenbeladung einer kinetischen Kontrolle unterliegt.

Durch die Berechnung der Titrationskurve konnte gezeigt werden, dass bei dem pH-Wert der optimalen Immobilisierung (pH 7) die Gesamtladung des Proteins negativ ist. Weitere Elektrostatik-Berechnungen zeigten, dass sich die Ladungsverteilung über die Proteinoberfläche bei diesem pH-Wert über große negative *patches* und einem kleinen positiven *patch* erstreckt. Der so genannte Nullpunkt von Aluminium-Silikat liegt etwa

bei pH 3 (O'Connor et al., 2006). Es wurde bereits gezeigt, dass geladene Proteine mit einer einhergehenden Änderung der Orientierung auch relativ gut auf gleichgeladenen Oberflächen bindet, sofern es einen entgegengesetzt geladenen *patch* enthält (Noh et al., 2008). Dadurch kann die erhöhte Immobilisierungseffizienz bei pH 7 erklärt werden. Außerdem konnte durch die Berechnungen gezeigt werden, dass das Protein auf der Matrix bei diesem pH-Wert die günstigste Orientierung einnimmt, da der positive *patch* sich auf der Reduktase-Interaktions-Seite befindet und so der Zugang zum aktiven Zentrum frei gehalten wird.

5.2 Laccasen

5.2.1 Die Laccase Engineering Database: Ein Klassifikations- und Analysesystem für Laccasen und verwandte Multikupferoxidasen

(siehe: *The Laccase Engineering Database: A Classification and Analysis System for Laccases and Related Multicopper Oxidases*, Seite 106)

Laccasen gehören zu der Familie der Multikupferoxidasen (*multicopper oxidases*, MCO). Funktionelle Studien haben gezeigt, dass die Substratbindung und -oxidation an einem Typ1-Kupferion (T1-Zentrum) und die Sauerstoffreduktion an einem trinukleären Kupfercluster, bestehend aus einem Typ2- und zwei Typ3-Kupferionen (T2/T3-Zentrum), erfolgt (Solomon et al., 1996). Multiple Sequenz- und Strukturanalysen ergaben vier Sequenzmuster (*pattern*), L1, M2, L3 und M4, die die Kupferbindestellen definieren. Die Signaturmuster M2 und M4 wurden als spezifische Muster für MCOs (Messerschmidt und Huber, 1990; Solomon et al., 2001) ausgelegt, während L1 und L3 als laccasespezifisch definiert wurden (Kumar et al., 2003). MCOs katalysieren die Oxidation einer Reihe, hauptsächlich phenolischer, Substrate. Insbesondere Laccasen, die die größte Subklasse der MCOs bilden, spielen wichtige Rollen in vielen zellulären und mikrobiellen Prozessen und sind auf Grund ihrer katalytischen Eigenschaften in vielen biotechnologischen Anwendungen von großer Bedeutung. Leider sind sie in ihrer Selektivität und ihrem Redox-Potential oft eingeschränkt. Ein besseres Verständnis des Zusammenhangs zwischen Sequenz, Struktur und Funktion von Laccasen und anderen homologen Proteinen der Familie der Multikupferoxidasen ist von großem Vorteil bei dem Entwurf von Proteinen mit verbesserten Eigenschaften. Daher wurde die *Laccase Engineering Database*

(*LccED*) innerhalb des etablierten *Data Warehouse* Systems *DWARF* (Fischer et al., 2006) implementiert.

Basierend auf früheren phylogenetischen Studien (Hoegger et al., 2006) wurden Superfamilien für (A) Basidiomycete-Laccasen, (B) Ascomycete-Laccasen, (C) Insekten-Laccasen, (D) pigmentbildende MCOs aus Pilzen, (E) pilzliche Ferroxidasen, (F) pilzliche und pflanzliche Ascorbatoxidasen, (G) pflanzliche laccase-ähnliche MCOs und bakterielle Laccasen, die weiter unterteilt wurden in (H) Kupfer-Resistenzproteine (CopA), (I) Bilirubinoxidasen, und (J) Proteine des Kupferflußsystems (CueO), erstellt. Die auf diese Weise erstellten Familien wurden mit nach Hoegger et al. (2006) eingeteilten 361 Sequenzen befüllt. Informationen zu ihnen wurden aus der *GenBank* (Benson et al., 2008) extrahiert. Diese Sequenzen dienten damit als Startsequenzen für die jeweilige Familie und wurden durch Homologie-Suchen mit *BLAST* (Altschul et al., 1997) um homologe Sequenzen erweitert. Die Unterteilung in homologe Familien erfolgte basierend auf phylogenetischen Analysen. Weiterhin wurden Strukturmonomere der *PDB* (Berman et al., 2002) entnommen und die *DSSP*-Sekundärstrukturinformationen (Kabsch und Sander, 1983) berechnet und innerhalb der Datenbank annotiert. Für alle Familien wurden mittels *CLUSTAL W* (Thompson et al., 1994) Multisequenzalignments und phylogenetische Bäume erstellt.

Die so erstellten wohldefinierten Proteinfamilien wurden im Hinblick auf Sequenzähnlichkeit, Herkunftsorganismus und die *pattern*, die die Kupferbindestellen charakterisieren, untersucht. Die Verteilung der Organismen über die Familien entsprach im Allgemeinen der ursprünglichen Zuweisung nach Hoegger et al. (2006). Da die Zuweisung homologer Proteine über Sequenzähnlichkeit erfolgt, war zu erwarten, dass sich auch Proteine aus verschiedenen Organismen in denselben Familien befinden, sogar Proteine verschiedener Lebensformen. Interessanterweise ergab eine automatische Mustersuche innerhalb der *LccED*, dass nur ca. 9 % aller MCOs die M2 und M4 enthalten und nur ca. 8 % aller Laccasen alle vier *pattern* enthalten. Eine manuelle Validierung der Multisequenzalignments ergab eine geringe Sensitivität der *pattern*, da sie eine hohe Anzahl an falsch-negativen Ergebnissen produzierten. Die Alignments wiesen eine sehr hohe Sequenzähnlichkeit in den Bereichen der vier *pattern* auf, doch ist die Definition aussagekräftiger *pattern* auf der Basis von nur 100 Sequenzen (Kumar et al., 2003) in einer Proteinfamilie einer solchen Größe und Diversität wahrscheinlich nicht möglich. Eine Erweiterung der *pattern* würde ihre Sensitivität steigern, jedoch würde damit unter Einbuße der Genauigkeit auch

die Zahl der falsch-positiven Ergebnisse steigen. Daher wurden für jede Familie der *LccED* familienpezifische *HMM*-Profile generiert, durch die Laccasen und andere MCOs charakterisiert werden können.

Die auf <http://www.lcced.uni-stuttgart.de> veröffentlichte Version der *LccED* enthält 2274 Proteine. Für 14 verschiedene Proteine aus 6 verschiedenen Familien sind Kristallstrukturen abgelegt. Die Proteineinträge sind unterteilt in 55 homologe Familien und 10 vordefinierte Superfamilien, die als Multisequenzalignments repräsentiert werden können. In den Alignments sind funktionale Aminosäuren farbig annotiert. Für jede Familie ist der zugehörige phylogenetische Baum und ein *HMM*-Profil abgelegt. Um unbekannte oder neue Sequenzen zuzuordnen, kann *online* auf der *LccED* eine *BLAST*-Suche (Altschul et al., 1997) durchgeführt werden. Der Inhalt der *LccED* wird in regelmäßigen Abständen automatisch aktualisiert und gewartet. Die gesamten Informationen der *LccED* können von dieser Seite heruntergeladen werden. Die *LccED* ist die erste *online*-Datenbank, die Informationen über Sequenzen, Strukturen, Alignments und Annotationen von Laccasen und ihrer homologen Proteine integriert. Durch die systematische Klassifikation der Proteine erlaubt sie die Zuweisung neuer Proteine. Basierend auf der *LccED* konnte eine umfassende Analyse aller Mitglieder im Rahmen der gesamten wohldefinierten Proteinfamilie durchgeführt werden, um den Zusammenhang zwischen Sequenzähnlichkeit, Organismen und Funktion zu verstehen. Außerdem konnten die an der Kupferbindung beteiligten Aminosäurereste identifiziert und annotiert werden.

Insgesamt ist mit der *LccED* ein verlässliches Werkzeug für die Analyse der Sequenz-Struktur-Funktionsbeziehung von Laccasen und verwandten Multikupferoxidasen im Kontext ihrer gesamten Proteinfamilie entstanden. Wie bereits für verschiedene Proteinfamilien, wie zum Beispiel Lipasen (Fischer und Pleiss, 2003), PHA-Depolymerasen (Knoll et al., 2009) und Cytochrom P450-Monooxygenasen (Sirim et al., 2009) gezeigt wurde, sind systematisch eingeteilte Proteinfamiliendatenbanken mit integrierten Analysewerkzeugen von großem Vorteil beim Design von Proteinen mit verbesserten Eigenschaften (Seifert und Pleiss, 2008; Seifert et al., 2009). Somit wurde durch die *LccED* ein vielversprechendes Werkzeug für die Optimierung von Enzymen durch *protein engineering* realisiert.

6 Publikationsmanuskripte und Publikationen in englischer Sprache

1. The Cytochrome P450 Engineering Database: A Navigation and Prediction Tool for the Cytochrome P450 Protein Family, Seite 60.
2. The Cytochrome P450 Engineering Database: Integration of Biochemical Properties, Seite 63.
3. The Modular Structure of Cytochrome P450 Monooxygenases, Seite 71.
4. Immobilization of P450 BM-3 on mesoporous molecular sieves, Seite 97.
5. The Laccase Engineering Database, Seite 106.

6.1 The Cytochrome P450 Engineering Database: A Navigation and Prediction Tool for the Cytochrome P450 Protein Family

Erschienen in *Bioinformatics* **23**: 2015-2017

Fischer, M., Knoll, M., Sirim, D., Wagner, F., Funke, S., Pleiss, J., 2007. The Cytochrome P450 Engineering Database: A Navigation and Prediction Tool for the Cytochrome P450 Protein Family.

Sequence analysis

The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family

Markus Fischer¹, Michael Knoll², Demet Sirim², Florian Wagner², Sonja Funke² and Juergen Pleiss^{2,*}

¹Department of Biochemistry & Molecular Biophysics, Columbia University, 1130 St. Nicholas Ave, New York, NY 10032, USA and ²Institute of Technical Biochemistry, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany

Received on March 20, 2007; revised on May 7, 2007; accepted on May 10, 2007

Advance Access publication May 17, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: The Cytochrome P450 Engineering Database (CYPED) has been designed to serve as a tool for a comprehensive and systematic comparison of protein sequences and structures within the vast and diverse family of cytochrome P450 monooxygenases (CYPs). The CYPED currently integrates sequence and structure data of 3911 and 25 proteins, respectively. Proteins are grouped into homologous families and superfamilies according to Nelson's classification. Nonclassified CYP sequences are assigned by similarity. Functionally relevant residues are annotated. The web accessible version contains multisequence alignments, phylogenetic trees and HMM profiles. The CYPED is regularly updated and supplies all data for download. Thus, it provides a valuable data source for phylogenetic analysis, investigation of sequence–function relationships and the design of CYPs with improved biochemical properties.

Abbreviations: Cytochrome P450 Engineering Database, CYPED; cytochrome P450 monooxygenase, CYP; Hidden Markov Model, HMM.

Availability: www.cyped.uni-stuttgart.de

Contact: Juergen.Pleiss@itb.uni-stuttgart.de

(<http://www.imm.ki.se/CYPalleles/>), the Directory of P450-containing Systems (<http://www.icgeb.trieste.it/>), the P450 Knowledgebase (<http://cpd.ibmh.msk.su/>) (Lisitsa *et al.*, 2001), the Arabidopsis P450 database (<http://www.p450.kvl.dk/>), the Insect P450 Site (<http://p450.antibes.inra.fr/>) or the P450s in PROMISE (<http://metallo.scripps.edu/PROMISE/P450.html>), a common data structure enabling the integration of available information on protein sequence and structure is still lacking. Therefore the Cytochrome P450 Engineering Database (CYPED) was implemented using the data warehouse system DWARF (Fischer *et al.*, 2006). The underlying data model assists the systematic analysis of the relationship of sequence, structure, and function of this vast and highly diverse protein family. The CYPED is the first cytochrome P450 data resource that combines information on sequences, sequence alignments, annotation and structures of CYPs. For data retrieval sequence, structure and annotation information is extracted from GenBank (Benson *et al.*, 2003) and PDB (<http://www.pdb.org/>). Functional annotation information is extended by an automated annotation transfer and was manually validated and enriched. Besides, the online accessible version, which is publicly available, supports the classification of unknown sequences by performing a BLAST search against the CYPED or by alignment to family-specific HMM profiles.

1 INTRODUCTION

Cytochrome P450 monooxygenases (CYPs) are heme containing enzymes that metabolize physiologically important compounds in many species of microorganisms, plants, animals and humans. CYPs catalyse the oxidation of a wide range of endogenous compounds in biosynthetic and biodegradation pathways, as well as xenobiotics such as drugs and environmental contaminants (Montellano, 1995). Therefore, understanding the substrate specificities of human CYPs is crucial for successful drug development (Raucy *et al.*, 2001). Although there are already numerous resources dedicated to CYPs like the Cytochrome P450 Homepage (<http://drnelson.utmem.edu/CytochromeP450.html>), the Homepage of the Human Cytochrome P450 (CYP) Allele Nomenclature Committee

2 DEVELOPMENT AND CONSTRUCTION

Seed sequences of CYPs were extracted from the Cytochrome P450 Homepage (Nelson *et al.*, 2002) and assigned to homologous families and superfamilies according to the Nelson classification scheme (Nelson, 2006). For each seed sequence, a BLAST search (Altschul *et al.*, 1997) was performed in the non-redundant sequence database at GenBank (Benson *et al.*, 2003) with a low E-value ($E = 10^{-100}$) to prevent overlapping hits among different superfamilies. For each hit, information on sequence, position-specific annotations, functional descriptions and the source organism was extracted and loaded by an automated retrieval system into an in-house developed relational database system (Fischer and Pleiss, 2003). New protein entries are assigned to homologous families and superfamilies according to their

*To whom correspondence should be addressed.

sequence similarity. The parameters are chosen as specified by Nelson. Proteins sharing a sequence identity of ≥ 40 or $\geq 55\%$ are members of the same superfamily or homologous family, respectively. About 2% of sequences were individually assigned to a family according to the recommendations of the P450 Nomenclature Committee. This procedure was applied to sequences which do not share a identity of $\geq 40\%$ with members of a homologous family but still belong to this family by definition. About 30% of the proteins within the database have not been classified by the nomenclature committee yet, and thus are assigned to the corresponding family by sequence similarity and named as 'homologous protein of family X (by similarity)'. They will be reassigned automatically during a database update in case of the classification information changes. Therefore, in contrast to existing P450 resources, the CYPED includes additional information which is expected to deepen the understanding in sequence-structure-function relationship of the CYP protein family and to apply the new gained knowledge to the design of improved CYPs.

CYP sequences that originate from the same organism and share a sequence identity of at least 98% are assigned to a single protein entry. For each protein entry the longest sequence was defined as reference sequence of the respective protein. For GenBank entries representing protein structures, monomers were extracted from the ExpDB database (Schwede *et al.*, 2000) and deposited as structure entries. Secondary structure information was calculated using DSSP (Kabsch and Sander, 1983), stored and annotated. To improve consistency and quality of the data, the classification into families and superfamilies for those protein entries that have not been classified was validated by performing multisequence alignments and a phylogenetic analysis. Multisequence alignment was also used to enrich annotation information and to control annotation quality. It was assumed that conserved sequence motifs should align for each superfamily. Therefore annotation information was transferred from one sequence to all other sequences in an alignment if the respective residues were conserved.

The CYPED is updated regularly. By an automated Perl script (Fischer *et al.*, 2006) new sequences are retrieved. New annotation information at GenBank for existing sequence entries and structure information is updated as well. Additionally, the update script takes care of changes of classification information.

3 DATA CONTENT AND ANNOTATED MULTISEQUENCE ALIGNMENTS

The CYPED contains sequence data on 3911 proteins. For 25 proteins of 20 different homologous families crystal structures are deposited. Since the CYPED provides a protein analysis tool to investigate the relationship between protein sequences, structures and their function, the data content is limited to sequences, structures and annotation information on amino acid level. The protein entries are assigned to 1111 homologous families, which are grouped into 531 superfamilies. Superfamilies and homologous families are represented as

multisequence alignments generated by CLUSTALW (Thompson *et al.*, 1994).

Common CYP motifs not present in GenBank entries were extracted from literature and annotated within CYPED entries. These motifs are the proline-rich region at the N terminus of CYPs (Kemper, 2004), the motif at the C-terminal end of helix K, the AGXXT motif present in helix I, and the functionally essential cysteine (Mestres, 2005). Amino acids linked to functional annotations are coloured within the alignments, and the residue number and further information is displayed upon moving the cursor over the respective amino acid. For each alignment, each column is coloured by the amino acid conservation score as calculated by PLOTCON (Rice *et al.*, 2000). For each homologous family and superfamily family-specific HMM profiles (<http://hmmer.janelia.org/>) are supplied.

4 WEB ACCESSIBILITY

The CYPED is accessible at <http://www.cyped.uni-stuttgart.de> by any JavaScript capable WWW browser. It can be browsed by superfamilies and homologous families, organisms, protein structures and the systematic CYP nomenclature. Protein names, source organisms, identifier codes and links to the corresponding GenBank entries are presented as tables. Protein sequences and trees can either be displayed with their accession codes as identifiers or their systematic names and source organisms. The hits are linked to the respective sequence entry, superfamily and homologous family. This functionality can also be applied to classify unknown sequences by performing a BLAST search on the web interface against the CYPED or by alignment to family-specific HMM profiles, which can be downloaded. All multisequence alignments, phylogenetic trees and structural monomers have been pre-calculated, and can be visualized or downloaded from this web site, too. Additionally, an archive can be downloaded comprising sequences, structures, alignments and phylogenetic trees grouped by families, and a formatted text file listing all protein information.

ACKNOWLEDGEMENTS

We acknowledge valuable contributions by Michael Krahn. This work was supported by the German Federal Ministry of Education and Research (project PTJ 0313080) and the German Research Foundation (SFB 706). We further are grateful for the unknown reviewer's comments which contributed greatly to the improvement of data content and the user interface of the CYPED. Funding to pay the Open Access publication charges was provided by the German Research Foundation.

Conflict of Interest: none declared.

REFERENCES

Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

- Benson,D.A. *et al.* (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
- Fischer,M. and Pleiss,J. (2003) The Lipase Engineering Database: a navigation and analysis tool for protein families. *Nucleic Acids Res.*, **31**, 319–321.
- Fischer,M. *et al.* (2006) DWARF – a data warehouse system for analyzing protein families. *BMC Bioinformatics*, **7**, 495.
- Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
- Kemper,B. (2004) Structural basis for the role in protein folding of conserved proline-rich regions in cytochromes P450. *Toxicol Appl. Pharmacol.*, **199**, 305–315.
- Lisitsa,A.V. *et al.* (2001) Cytochrome P450 database. *SAR QSAR Environ Res.*, **12**, 359–366.
- Mestres,J. (2005) Structure conservation in cytochromes P450. *Proteins*, **58**, 596–609.
- Montellano,O.d. (1995) Cytochrome P450: structure, mechanism and biochemistry. New York, Plenum Press.
- Nelson,D.R. (2006) Cytochrome P450 nomenclature, 2004. *Methods Mol. Biol.*, **320**, 1–10.
- Nelson,D.R. *et al.* (2002) Mining databases for cytochrome P450 genes. *Methods Enzymol.*, **357**, 3–15.
- Raucy,J.L. and Allen,S.W. (2001) Recent advances in P450 research. *Pharmacogenomics J*, **1**, 178–186.
- Rice,P. *et al.* (2000) EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 133–154.
- Schwede,T. *et al.* (2000) Protein structure computing in the genomic era. *Res Microbiol.*, **151**, 107–112.
- Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

6.2 The Cytochrome P450 Engineering Database: Integration of Biochemical Properties

Erschienen in *BMC Biochemistry* **10**: 27

Sirim, D., Wagner, F., Lisitsa, A., Pleiss, J., 2009. The Cytochrome P450 Engineering Database: Integration of Biochemical Properties.

6.2.1 Abstract

Background

Cytochrome P450 monooxygenases (CYPs) form a vast and diverse enzyme class of particular interest in drug development and a high biotechnological potential. Although very diverse in sequence, they share a common structural fold. For the comprehensive and systematic comparison of protein sequences and structures the Cytochrome P450 Engineering Database (*CYPED*) was established. It was built up based on an extensible data model that enables its functions readily enhanced.

Description

The new version of the *CYPED* contains information on sequences and structures of 8613 and 47 proteins, respectively, which strictly follow Nelson's classification rules for homologous families and superfamilies. To gain biochemical information on substrates and inhibitors, the *CYPED* was linked to the Cytochrome P450 Knowledgebase (*CPK*). To overcome differences in the data model and inconsistencies in the content of *CYPED* and *CPK*, a metric was established based on sequence similarity to link protein sequences as primary keys. In addition, the annotation of structurally and functionally relevant residues was extended by a reliable prediction of conserved secondary structure elements and by information on the effect of single nucleotide polymorphisms.

Conclusions

The online accessible version of the *CYPED* at <http://www.cyped.uni-stuttgart.de> provides a valuable tool for the analysis of sequences, structures and their relationships to biochemical properties.

6.2.2 Background

Cytochrome P450 monooxygenases (CYPs) constitute one of the largest superfamilies of enzymes, spread widely among species of microorganisms, plants, animals, and humans. Since they catalyze the oxidation of a wide range of endogenous compounds in biosynthetic and biodegradation pathways, as well as xenobiotics such as drugs and environmental contaminants (Montellano, 1995), an understanding of the substrate specificities of CYPs is crucial for successful drug development and biotechnological applications (Raucy and Allen, 2001). CYPs require interaction with a reductase, either as separate protein or as fusion protein (McLean et al., 2005).

We established the *CYPED* (Fischer et al., 2007) as a tool for a comprehensive and systematic comparison of CYP sequences and structures, which share only a very low percentage of sequence identity between the superfamilies (Graham and Peterson, 1999). For this purpose seed sequences have been extracted from the Cytochrome P450 Homepage (Nelson, 2002), incorporated in our in-house data warehouse system *DWARF* (Fischer et al., 2006), updated by a BLAST (Altschul et al., 1997) search and assigned to homologous families and superfamilies according to the recommended classification scheme (Nelson, 2006). Since the publication of the *CYPED*, it was applied to identify selectivity and specificity determining residues (Seifert and Pleiss, 2008), to adjust CYP families in the Fungal Cytochrome P450 Database (Park et al., 2008) and served as a template to design other protein family databases (Knoll et al., 2009; Knoll and Pleiss, 2008). The amount of available CYP sequences and structures almost doubled. Therefore, besides integrating new sequences and structures, we extended the *CYPED* by biochemical properties, and by adding new functionalities:

- Information on P450-catalyzed reactions, substrate preferences, induction and inhibition is made available by the *CPK* (Lisitsa et al., 2001). Since the protein identifiers of the two databases *CYPED* and *CPK* could not be related un-ambiguously, an algorithm which uses a metric based on sequence similarities was developed to link protein entries.
- Information on single-nucleotide polymorphism in human CYP sequences was extracted from the CYPallele homepage (Oscarson and Ingelman-Sundberg, 2002).
- CYPs share highly conserved secondary structure elements (Mestres, 2005). Therefore it was possible to reliably predict these elements from sequence and annotate them in the *CYPED*.

6.2.3 Construction and content

Database establishment

Homologous families and superfamilies were named according to the Cytochrome P450 Homepage (Nelson, 2002) and filled with consistently named CYP sequences from the first version of the *CYPED*. Thus, seed sequences for almost 400 superfamilies could be identified. Positions 1-499 were annotated as P450-domain to avoid loading reductases into the *CYPED* while updating fusion enzymes. For each seed sequence a BLAST search (Altschul et al., 1997) was performed in the non-redundant sequence database at NCBI (<http://www.ncbi.nlm.nih.gov>) with an E-value of 10^{-100} . For each hit, information on sequence, position specific annotations, functional descriptions, and the source organism was extracted and loaded by an automated retrieval system into an in-house developed relational database system (Fischer et al., 2006). In 28 % of the entries the correct CYP name according to the Nelson's classification (Nelson, 2006) was provided in the NCBI database entry. In 1 % the name was in contrast to sequence similarity, and therefore the protein was re-assigned. 1 % of the proteins had a name which does not exist according the Nelson scheme and therefore were assigned to the most similar existing family. Entries which were lacking information on the CYP name were assigned to a family by sequence similarity. Thus 64 % of the proteins could be assigned which have not been classified yet. All sequences which were assigned only based on sequence similarity were labeled by "homologous protein of family X (BY SIMILARITY)". 218 proteins without CYP name information and no sequence similarity to existing families, as well as 279 protein fragments were discarded. Following this procedure the entries of the *CYPED* are consistent with the recommendations of the nomenclature committee.

Sequence entries that originate from the same organism and share a sequence identity of at least 98 % are assigned to a single protein entry. For proteins with multiple sequence entries, the longest sequence was defined as reference sequence of the respective protein. Protein structures were downloaded from the Protein Data Bank (PDB) (Berman et al., 2000) and stored as structural monomers. Secondary structure information was calculated using DSSP (Kabsch and Sander, 1983). Information on structurally or functionally relevant residues was extracted from the GenBank and annotated in the *CYPED*.

New features and functionalities

The current version of the *CYPED* also provides a feature page for each protein entry where the sequence is displayed and annotations are highlighted. A newly developed dynamic web interface directly incorporates changes in the database.

The integration of protein databases is based on a common, unique key. A database-independent attribute of a protein which can be applied like a primary key is the protein sequence itself. While a primary key has to be specific, sequences can slightly vary although they might belong to the same protein entry. To overcome this problem, an algorithm was implemented (figure 6.2.1) which allows the direct use of the sequences as primary keys without the requirement of being completely identical which was termed a metric primary key. For each *CYPED* entry, a BLAST search against the *CPK* database was performed. The BLAST hits were ranked by E-value and a global pairwise alignment was performed (Needleman and Wunsch, 1970). The *CPK* entries with a sequence identity of more than 90 % are displayed on the protein feature page, linking to the corresponding *CYPED* sequence to the respective entries in the *CPK*. Thus, the sequences can be applied as common attribute of protein entries and serve as primary keys.

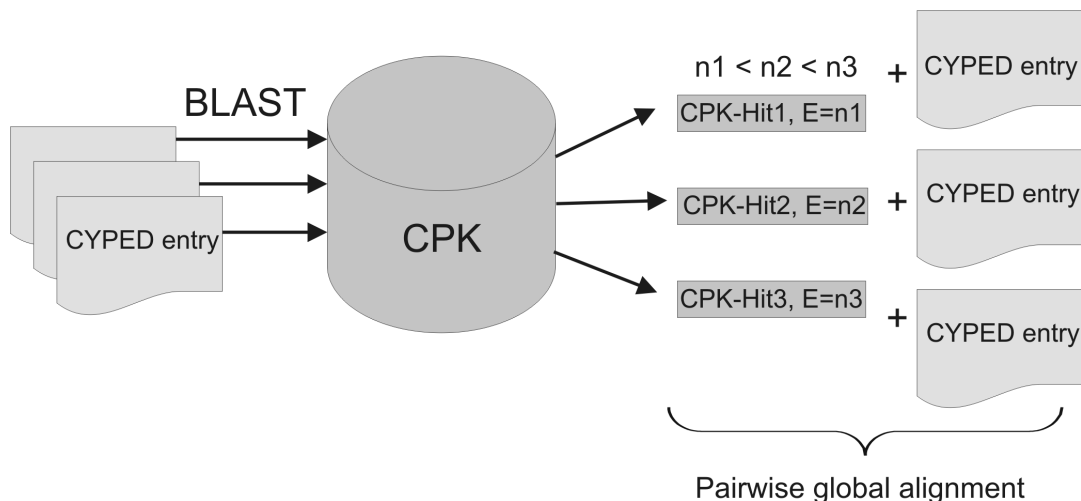


Figure 6.2.1: *CYPED* - *CPK* integration pipeline.

Identification and assignment algorithm of the *CYPED* proteins and the corresponding *CPK* entries. The steps of the algorithm involve a BLAST search of each of the *CYPED* entries against the *CPK*, a ranking of the hits by E-value and a final pairwise alignment of the original *CYPED* entry with the corresponding *CPK*-hits to obtain the percentage identity.

For all sequence entries the conserved secondary structures were predicted by a structure-based HMM profile which was embedded in an automated annotation program, stored as annotations in the *DWARF*-system and are displayed on the protein feature pages and within the multisequence alignments.

Information on human CYP alleles are extracted from the "Home Page of the Human Cytochrome P450 (CYP) Allele Nomenclature Committee" (Oscarson and Ingelman-Sundberg, 2002) and stored in tables designated for this purpose in the database. The mutations and their effect, whether the enzymes lack of activity or gained increased activity, are listed on the protein feature page.

Contents

The *CYPED* contains 11193 sequence entries for 8613 protein entries. The proteins have been assigned to 249 superfamilies and 619 homologous families. Structure information for 47 different proteins which originate from 36 different homologous families was extracted from 228 PDB entries.

In total, 3575 *CYPED* proteins matched the respective *CPK* entries with a sequence identity of more than 90 %. These matches provided the links to 3257 different compounds (1699 substrates, 723 inducers and 1227 inhibitors). This information has been extracted from more than 10000 research papers cited in PubMed (Lisitsa et al., 2001).

For each family, a multisequence alignment and a phylogenetic tree were generated by CLUSTALW (Thompson et al., 1994). The annotated version is colour-coded and highlights functionally relevant sites and the predicted secondary structure. For each alignment, the degree of conservation of each column is indicated on the bottom as a coloured chart as calculated by PLOTCON (Rice et al., 2000). For each homologous family and superfamily, family specific HMM profiles (<http://hmmer.janelia.org/>) are supplied.

6.2.4 Utility and discussion

The online version of the *CYPED* can be browsed by families, source organisms, or structures. Pre-calculated multisequence alignments and structural monomers are displayed and can be downloaded. Phylogenetic trees are visualized by the program PHYLODEN-

DRON (<http://iubio.bio.indiana.edu/treeapp/>). Sequences in the alignments and trees can either be displayed with their accession codes or their systematic names and source organisms as identifiers and they are linked to the respective GenBank entry. For each superfamily and homologous family, family-specific HMM profiles are provided, which can be applied for the classification and the identification of new CYP sequences. Besides, the website provides a local BLAST interface where a homology search can be performed against the *CYPED*.

For the 3575 protein entries in the *CYPED* that match *CPK* entries, the interface to the *CPK* provides biochemical information on substrates, inhibitors, and inducers. In the *CPK*, effectors of CYPs are separated into drugs and non-drugs. Special tags show the cases, when the same compound was reported as substrate and inducer, or when experimental results negated the CYP activity towards a certain chemical species. Through the hyperlink the user is also provided with the information on the PubMed references reporting the relationship between a CYP isoform and low-molecular effector.

6.2.5 Conclusions

The Cytochrome P450 Engineering Database (*CYPED*) provides a collection of tools for classification and analysis of the vast and diverse family of cytochrome P450 monooxygenases. To gain a better understanding in biochemical properties and sequence-structure-function relationships, the features of the *CYPED* were extended by integrating biochemical information from the *CPK*. Thus, the *CYPED* has become a valuable tool to navigate in sequence space and to analyze sequence-structure-function relationships.

6.2.6 Availability and requirements

All sequences, multisequence alignments, phylogenetic trees, and HMM profiles of the Cytochrome P450 Engineering Database (*CYPED*) are accessible via a web interface at <http://www.cyped.uni-stuttgart.de>. Additionally, all data is supplied for download.

6.2.7 List of abbreviations

CYP: Cytochrome P450 monooxygenase; CYPED: Cytochrome P450 Engineering Database; CPK: Cytochrome P450 Knowledgebase; DWARF: Data Warehouse System for Analyzing Protein Families; BLAST: Basic Local Alignment Search Tool; HMM: Hidden Markov model; DSSP: Define Secondary Structure of Proteins.

6.2.8 Authors' contributions

DS established the database and wrote the manuscript. FW designed and implemented the integration algorithm and generated the web interface. AL provided the *CPK* data and contributed to the data integration and to the manuscript. JP supervised the project and finalized the manuscript.

6.2.9 Acknowledgements

This work was supported by the German Research Foundation (SFB 706).

6.2.10 References

- Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25 (17), 3389–3402, 1997.
- Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* 28, 235–242, 2000.
- Fischer, M., Knoll, M., Sirim, D., Wagner, F., Funke, S., Pleiss, J., The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. *Bioinformatics* 23 (15), 2015–2017, 2007.
- Fischer, M., Thai, Q. K., Grieb, M., Pleiss, J., DWARF—a data warehouse system for analyzing protein families. *BMC Bioinformatics* 7, 495, 2006.
- Graham, S. E., Peterson, J. A., How similar are P450s and what can their differences teach us? *Arch Biochem Biophys* 369 (1), 24–29, 1999.
- Kabsch, W., Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12), 2577–2637, 1983.
- Knoll, M., Hamm, T. M., Wagner, F., Martinez, V., Pleiss, J., The PHA Depolymerase Engineering Database: A systematic analysis tool for the diverse family of polyhydroxyalkanoate (PHA) depolymerases. *BMC Bioinformatics* 10, 89, 2009.
- Knoll, M., Pleiss, J., The Medium-Chain Dehydrogenase/reductase Engineering Database: a systematic analysis of a diverse protein family to understand sequence-structure-function relationship. *Protein Sci* 17 (10), 1689–1697, 2008.

- Lisitsa, A. V., Gusev, S. A., Karuzina, I. I., Archakov, A. I., Koymans, L., Cytochrome P450 database. SAR QSAR Environ Res 12 (4), 359–366, 2001.
- McLean, K. J., Sabri, M., Marshall, K. R., Lawson, R. J., Lewis, D. G., Clift, D., Balding, P. R., Dunford, A. J., Warman, A. J., McVey, J. P., Quinn, A. M., Sutcliffe, M. J., Scrutton, N. S., Munro, A. W., Biodiversity of cytochrome P450 redox systems. Biochem Soc Trans 33 (Pt 4), 796–801, 2005.
- Mestres, J., Structure conservation in cytochromes P450. Proteins 58 (3), 596–609, 2005.
- Montellano, O. d., Cytochrome P450: structure, mechanism and biochemistry. New York, Plenum Press, 1995.
- Needleman, S. B., Wunsch, C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48 (3), 443–453, 1970.
- Nelson, D. R., Mining databases for cytochrome P450 genes. Methods Enzymol 357, 3–15, 2002.
- Nelson, D. R., Cytochrome P450 nomenclature, 2004. Methods Mol Biol. 320, 1–10, 2006.
- Oscarson, M., Ingelman-Sundberg, M., CYPalleles: a web page for nomenclature of human cytochrome P450 alleles. Drug Metab Pharmacokinet 17 (6), 491–495, 2002.
- Park, J., Lee, S., Choi, J., Ahn, K., Park, B., Park, J., Kang, S., Lee, Y.-H., Fungal cytochrome P450 database. BMC Genomics 9, 402, 2008.
- Raucy, J. L., Allen, S. W., Recent advances in P450 research. Pharmacogenomics J 1 (3), 178–186, 2001.
- Rice, P., Longden, I., Bleasby, A., EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16 (6), 276–277, 2000.
- Seifert, A., Pleiss, J., Identification of selectivity-determining residues in cytochrome P450 monooxygenases: A systematic analysis of the substrate recognition site 5. Proteins, 2008.
- Thompson, J. D., Higgins, D. G., Gibson, T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22 (22), 4673–4680, 1994.

6.3 The Modular Structure of Cytochrome P450 Monooxygenases

Zur Einreichung bei *BMC Structural Biology*

Sirim, D., Wagner, F., Widmann, M., Pleiss, J., The Modular Structure of Cytochrome P450 Monooxygenases.

6.3.1 Abstract

Background:

Cytochrome P450 monooxygenases (CYPs) form a vast and diverse family of highly variable sequences. They catalyze a wide variety of oxidative reactions and are therefore of great relevance in drug development and biotechnological applications. Despite their differences in sequence and substrate specificity, the structures of CYPs are highly similar. Although being in research focus for years, factors mediating selectivity, and activity remain vague.

Results:

This systematic comparison of CYPs based on the Cytochrome P450 Engineering Database (*CYPED*) involved sequence and structure analysis of more than 8000 sequences. 31 structures have been applied to generate a reliable structure-based HMM profile in order to predict structurally conserved regions. Therefore, it was possible to automatically transfer these modules on sequences without any secondary structure information, to analyze substrate interacting residues and to compare interaction sites with redox partners.

Conclusion:

Functionally relevant structural sites of CYPs were predicted. These regions involved in substrate binding as well as redox partner recognition and interaction were extensively analyzed in all sequences among the *CYPED*. The newly gained insights promise an improvement of engineered enzyme properties for potential biotechnological application. The annotated sequences are accessible on the current version of the *CYPED*. The prediction tool can be applied to any CYP sequence via the web interface at <http://www.cyped.uni-stuttgart.de/cgi-bin/strpred/dosecpred.pl>.

6.3.2 Background

Cytochrome P450 monooxygenases (CYPs) are a ubiquitous protein family, existing in all eukaryotes, most prokaryotes and Archae. These heme-containing enzymes catalyze the monooxygenation of a large variety of substrates (Montellano, 1995). CYPs have an essential function in drug metabolism, hence focussed in the pharmaceutical industry (Raucy and Allen, 2001). Besides, they are of great interest for synthetical application in biotechnology as versatile biocatalysts (Urlacher and Eiben, 2006). A profound knowledge in the factors mediating selectivity and activity of these proteins is a prerequisite in the development of CYPs with improved properties. Therefore, deeper insights in the relationships between sequence, structure and function are of great interest.

According to Nelson's classification (Nelson, 2006) CYPs are grouped into homologous families and superfamilies, predominantly based on sequence similarity. The sequence identity between proteins from different superfamilies is extremely low and may be less than 20 % (Graham and Peterson, 1999). Only three amino acids are totally conserved, among these the glutamic acid and the arginine of the ExxR-motif, which is involved in stabilizing the core and heme-binding (Hasemann et al., 1995), and the heme-binding cysteine. However, the increasing number of crystal structures shows that despite this unusual variability the overall structure is highly conserved: CYPs consists of structural conserved modules that are essential for structure and function, and of variable regions that mediate the individual biochemical properties. The defined conserved secondary structures are named α A-L and β 1-5 and could be identified in all CYP structures and make up the so called CYP-fold (de Graaf et al., 2005; Werck-Reichhart and Feyereisen, 2000; Peterson and Graham, 1998).

Most CYPs require interaction with a reductase to provide electrons, either as separate protein or as fusion protein. Depending on the nature of their electron transfer partner, CYPs are assigned to different classes. There are several proposed classification schemes which subdivide CYPs in up to nine classes (McLean et al., 2005; Munro et al., 2007; Hannemann et al., 2007). The most simple one discriminates between three classes of CYPs: class I, which comprises CYPs interacting with a cytochrome P450 reductase-type (CPR-type) FMN/FAD reductase, class II, which comprises CYPs interacting with other redox systems, and class III proteins which do not need a reductase for their reaction (Baudry et al., 2006). CYPs such as CYP 102A1 from *Bacillus megaterium* (P450 BM-3) are fusion proteins between a heme domain and a reductase and CYPs of

class III appear very rarely in nature (Munro et al., 2002). Therefore, in most CYPs the interaction with their appropriate redox partner is prerequisite for their reaction to occur. Many different CYP isoenzymes interact with only one reductase, and it is assumed that CYPs of the same class are similar in their reductase interaction sites (Bernhardt, 1996). It is expected that there are favorable electrostatic interactions between CYPs and their electron transfer partner (Wade et al., 2005). A crystal structure for a CYP-reductase-complex is yet not available. The electron transfer from the reductase to the heme domain often is slow and rate-limiting (Guengerich, 2002), and the interactions between the components of the electron transfer systems still remain unclear. A deeper understanding of the factors determining reductase interaction gained by the analysis of the reductase interaction sites of CYPs will assist in improving interactions and consequently be leading to significantly optimized enzymes for biocatalytic applications (Bernhardt, 2006).

Previous analysis of the structure conservation in CYPs showed that all CYPs have a well-conserved heme-binding structural core formed out of α D, α E, α I, and α L and α J and α K (Mestres, 2005). The β -bulge region which contains the thiolate heme ligand is referred to as Cys-pocket. Between α K and the Cys-pocket, a structurally conserved region is located, the so-called 'meander' loop. It is spanned by 7-10 amino acid residues and is supposed to play a role in heme binding and stabilization of the tertiary structure. The proposed reductase interaction face of CYPs mainly comprises the α J/ α J' and the insertion following the meander loop (Hasemann et al., 1995). Since the structures of all CYPs are highly similar, but differ in substrate specificity and their electron transfer partners, the different biochemical properties of CYPs are mediated by the diverse regions, which vary in both sequence and structure (Peterson and Graham, 1998). Six regions which are involved in recognition and binding of substrates and hence determine substrate specificity were previously described as SRSs (substrate recognition sites, Gotoh (1992)). SRS1 lies in the highly variable BC-loop region, SRS2 is located in the C-terminal end of α F, SRS3 and SRS4 are spanned by the N-terminal regions of α G and α I, β 1-4 houses SRS5 and β 4-1 SRS6. The α I, the BC-loop region and SRS5 are limiting the access of the substrate to the heme. In a systematic analysis of SRS5 in more than 6300 sequences, single substrate- and heme-interacting residues could be identified in this region (Seifert and Pleiss, 2008): Thus, a hotspot for regio- and stereoselectivity in one residue in SRS5 and one position in the BC-loop (F87), previously were reported as key residues in determination activity, regio- and stereoselectivity in P450 BM-3 (Li et al., 2008; Urlacher and Schmid, 2002; Urlacher et al., 2006). Combinations of variants of these

two positions were applied to design a minimal mutant library with improved selectivity (Seifert et al., 2009). Due to the high variability of the BC-loop, the identification of position 87 in P450 BM-3 in other CYPs, remains a challenge for sequences without structural information.

To serve as a tool for a comprehensive comparison of protein sequences and structures within the vast and diverse family of CYPs in order to transfer the newly gained insights among the CYP sequences, the Cytochrome P450 Engineering Database (*CYPED*) (Fischer et al., 2007) has been designed. In its current version 2.02 it contains 8614 sequences (Sirim et al., 2009). The highly similar structures have been compared in detail to identify the common core and to assign the variable regions. Therefore, a structural alignment was used as base to generate a reliable structure profile. With this profile all structurally conserved regions (SCR) could be predicted and annotated among all *CYPED* protein sequence entries, hence allowing a structural navigation in those sequences lacking structural information. Beyond this, the *CYPED* website provides an interface which makes the prediction of the SCRs possible for any CYP sequence.

6.3.3 Data

CYP Structures

A set of 31 PDB structures (Berman et al., 2002) was extracted from version 1.1 of the *CYPED* (Fischer et al., 2007) as listed in table 6.3.1. The selection includes 12 CYPs of class I, comprising CYPs which interact with a CPR-type FMN/FAD reductase. The structures in this class are predominantly of mammalian origin. The only exception is P450 BM-3 from *Bacillus megaterium*, which is a fusion enzyme, consisting of a P450 domain and a FMN/FAD reductase domain (Munro et al., 2007). Because of its structural closeness to P450 BM-3, the bacterial CYP175A1 isolated from the thermophilic *Thermus thermophilus* was also assigned to class I (Baudry et al., 2006). 16 structures of bacterial or fungal origin were assigned to class II, which interact with redox partners other than CPR. Crystal structures for class III CYPs are: CYP8A (human prostacyclin synthase), which accepts endoperoxides or hydroperoxides as substrates and does not require any electron-transfer partner or molecular oxygen (Chiang et al., 2006); CYP55A2 from *Fusarium oxisporum* and 152A1 from *Bacillus subtilis* (P450_{Bsβ}) are representants for CYPs which obtain electrons directly from NAD(P)H or catalyze a peroxide-dependent reaction. Eleven recently published CYP structures were used to validate the prediction.

Table 6.3.1: List of CYP structures analyzed in this work

Class I CYPs with structures (CPR-type)			
CYP	PDB-Code	Resolution [Å]	Organism
1A2	2HI4	1.95	<i>Homo sapiens</i>
2A6	1Z10	1.90	<i>Homo sapiens</i>
2A13	2P85	2.35	<i>Homo sapiens</i>
2B4	1SUO	1.90	<i>Oryctolagus cuniculus</i>
2C5	1N6B	2.30	<i>Oryctolagus cuniculus</i>
2C8	1PQ2	2.70	<i>Homo sapiens</i>
2C9	1OG2	2.60	<i>Homo sapiens</i>
2D6	2F9Q	3.00	<i>Homo sapiens</i>
2R1	2OJD	2.80	<i>Homo sapiens</i>
3A4	1TQN	2.05	<i>Homo sapiens</i>
102A1	1BU7	1.65	<i>Bacillus megaterium</i> (P450 BM-3)
175A1	1N97	1.80	<i>Thermus thermophilus</i>
Class II CYPs with structures			
CYP	PDB-Code	Resolution [Å]	Organism
51B1	1E9X	2.10	<i>Mycobacterium tuberculosis</i>
101D	2CPP	1.63	<i>Pseudomonas putida</i> (P450cam)
107A1	1OXA	2.10	<i>Saccharopolyspora erythraea</i>
107L1	2BVJ	2.10	<i>Streptomyces venezuelae</i>
108A	1CPT	2.30	<i>Pseudomonas sp.</i>
119	1IO7	1.50	<i>Sulfolobus solfataricus</i>
154A1	1ODO	1.85	<i>Streptomyces coelicolor</i>
154C1	1GWI	1.92	<i>Streptomyces coelicolor</i>
158A1	2DKK	1.97	<i>Streptomyces coelicolor</i>
158A2	1S1F	1.50	<i>Streptomyces coelicolor</i>
165B3	1LFK	1.70	<i>Amiclatopsis orientalis</i>
165C4	1UED	1.90	<i>Amiclatopsis orientalis</i>
167A1	1Q5D	1.93	<i>Polyangium cellulorum</i>
176A1	1T2B	1.70	<i>Citrobacter braakii</i>
199A2	2FR7	2.01	<i>Rhodospseudomonas palustris</i>
245A1	2Z3T	1.90	<i>Streptomyces sp.</i> TP-A0274
Class III CYPs with structures			
CYP	PDB-Code	Resolution [Å]	Organism
8A	2IAG	2.15	<i>Homo sapiens</i>
55A2	1CL6	1.70	<i>Fusarium oxysporum</i> (NO reductase)
152A1	1IZO	2.10	<i>Bacillus subtilis</i> (P450 _{Bsβ})

CYP Sequences

The analysis of CYP sequences and structures was performed based on the updated version 2.02 of the *CYPED* (Sirim et al., 2009). It integrates sequences of 8614 proteins. The proteins are organized into 249 superfamilies and 619 homologous families according to Nelson (Nelson, 2006). Reliable multisequence alignments are available for each family. The sequences are annotated by automatically extracted GenBank annotations (Benson et al., 2008), which were manually enriched. Secondary structure information is available as DSSP annotation within the multisequence alignments for those homologous families containing members with existing PDB structures.

6.3.4 Methods

Structure-based HMM profile

SCRs were determined by the generation of a structural alignment using STAMP (Russell and Barton, 1992). STAMP estimates the probability of structural equivalence of residues (Rossmann and Argos, 1976) and uses the Smith-Waterman algorithm (Smith and Waterman, 1981) to determine the best path through a matrix of numerical pairwise similarity values of corresponding sequence positions. This allows STAMP to calculate two measures of alignment confidence: P'_{ij} , a measure for residue equivalence and S_c , the STAMP score, which reflects overall alignment quality. A $S_c > 5.5$ implies a high degree of similarity of the considered structures. Stretches of residues having $P'_{ij} > 6.0$ imply regions of conserved secondary structure. To visualize secondary structure information on the alignment output, STAMP uses DSSP (Kabsch and Sander, 1983) outputs. Therefore in a first step DSSP was applied on the CYP structures to calculate secondary structure information. The resulting alignment was checked for correctly aligned secondary structures, ExxR motif and Cys-pocket. SCRs are extracted from the alignment and visualized (figure 6.3.1) on the structure from CYP102A1 (PDB: 1BU7) as reference structure using PyMOL (Delano, 2002). From the alignments HMM-profiles were generated using HMMER (<http://hmmer.janelia.org/>).

Structural analysis

Structural superpositions and visualization were generated using PyMOL (Delano, 2002). The analysis of the BC-loop region was performed by a superposition of all structures on P450 BM-3 (PDB: 1BU7). The visualization of reductase interaction sites RIS1 and RIS2 was generated by the superposition of the FMN-domains of P450 BM-3 (PDB:

1BVY) and the CPR-type FMN/FAD reductase from *Rattus norvegicus* (PDB: 3ES9) and structurally aligning the P450-domains of CYP2C9 from *Homo sapiens* (PDB: 1OG2) CYP101D from *Pseudomonas putida* (PDB: 2CPP) on the P450-domain of P450 BM-3.

Sequence analysis

For analysis of all CYP sequences, the *CYPED* and the *DWARF* system (Fischer et al., 2006) were applied. The data warehouse system *DWARF* is the in-house repository for the *CYPED* data and assists local analysis. Besides integrating sequences and structures of the well-defined protein family, it provides a set of bioinformatics tools for sequence and structure analysis. We took advantage of its modular and extensible architecture and designed a Perl program which implements an automated procedure that subsequently generates a structure-based alignment for every *CYPED* entry by mapping it on the structure-based HMM profile. The start and stop positions of each conserved secondary structure were identified within each alignment among the sequence of the structure of P450 BM-3 as reference. It was assumed that the corresponding positions of the query sequence can be found in the same columns of the alignment. Therefore, the absolute positions of the SCRs of each query sequence could be calculated. The positions were stored as annotations in the *CYPED* and are displayed among the multisequence alignments and on the feature page of each *CYPED* entry.

The same procedure as for the identification of the SCRs was applied to identify the specificity and regioselectivity determining position which corresponds F87 in P450 BM-3 in all sequences among the *CYPED*. Each *CYPED* sequence was mapped on the profile of the structural alignment and the phenylalanine at position 87 was determined in the sequence of the structure belonging to P450 BM-3. The corresponding residue in the query sequence was assumed in the same column and extracted for each sequence. The accuracy of this method was tested in a leave-one-out cross-validation (Picard and Cook, 1984) by generating 30 different structure-based HMM profiles and subsequently mapping the sequence of the left-out crystal structure on them.

An online version of the prediction tool was also integrated into the *CYPED* homepage. Since the method operates exclusively for sequences with CYP fold, input sequences first are checked for applicability by sequence homology via a BLAST (Altschul et al., 1997) query using an E-value of 10^{-100} . Structurally conserved regions are determined as described above.

6.3.5 Results

Structural core

From the simultaneous superposition of the 31 structures using STAMP, a multiple sequence alignment could be derived which resulted in 257 structurally equivalent residues out of 400-450 residues. The calculated average RMS deviation after fitting all structures by these 257 residues was 2.4 Å and their averaged sequence identity was 25 %. The overall STAMP alignment score S_c was 6.0 and is above the threshold for highly similar structures. Stretches of structurally equivalent residues ($P'_{ij} > 6.0$) are marked by black boxes in the structure-based sequence alignment (supplementary figure 6.3.7). The residues of the conserved core are organized into 19 SCRs that include at least partially all defined secondary structures α A-L and β 1-4. The SCRs extracted from the structural alignment were mapped on the reference structure P450 BM-3 from *Bacillus megaterium* (figure 6.3.1).

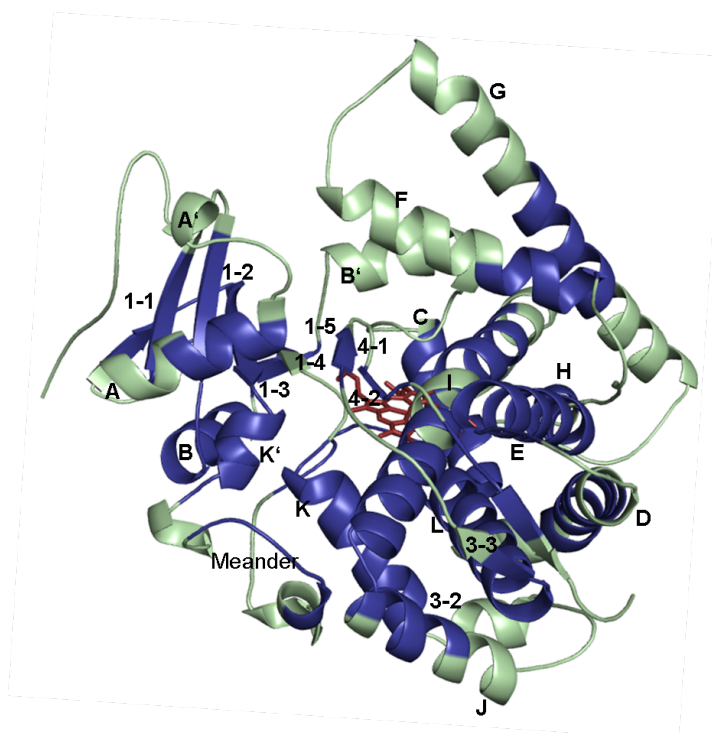


Figure 6.3.1: Conserved regions derived from STAMP alignment mapped on reference structure P450 BM-3 from *Bacillus megaterium* (PDB: 1BU7). The SCRs are highlighted in blue, whereas the variable regions are shown in green.

A topological overview of the conserved CYP structure illustrates the distribution of SCRs on the CYP structure (figure 6.3.2). Some SCRs are part of individual secondary structures, other SCRs include several secondary structure elements. Among these, SCR3

comprises β 1-2 and α B, SCR7 β 3-1 and α E. SCR11 is assembled by α I and α J and SCR13 by β 1-4 and β 2-1. β 2-2, β 1-3 and α K' together form SCR14 and the heme-binding Cys-pocket and α L together form SCR16. The structural alignment further revealed that the β -5 sheet which is not present in all CYP structures does not belong to the conserved parts of the CYP structures (Baudry et al., 2006). The variable termini of the secondary structure elements α F, α G, α I, β 1-4, β 4-1, and the BC-loop are surrounding the heme and house the residues defining the SRS regions 1-6.

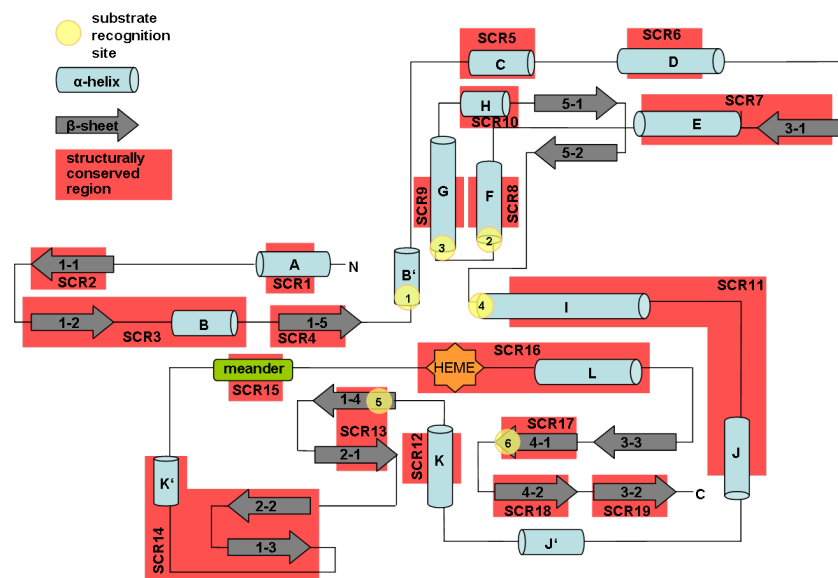


Figure 6.3.2: Conserved regions derived from STAMP alignment in a topological overview.

By applying the procedure on each *CYPED* sequence and mapping it on the HMM profile generated from the STAMP alignment, the SCRs could be identified and annotated in all sequence entries. The conserved secondary structures appear in the online version of the *CYPED* either within the annotated multisequence alignments or on the feature page of each protein entry. Its labelling appears in moving over the respective region. The results of the online prediction for any CYP sequence are displayed as colored and annotated regions and as a tabular output listing each conserved secondary structure and the corresponding start and stop position.

BC-loop

In P450 BM-3 the phenylalanine at position 87 is assumed to mediate selectivity and activity. Due to its proximity to the heme center, this residue has a strong evidence to be involved in substrate binding and to control substrate specificity and regioselectivity

(Seifert et al., 2009). Therefore, the identification of residues corresponding to this position would be beneficial in the design of CYPs with engineered properties. Since it is located in the SRS1 region of the highly variable BC-loop the identification of this position in enzymes without structural information is not possible merely by sequence alignment. However, a comprehensive analysis of the BC-loops in the structures analyzed in this work revealed that although being highly variable (figure 6.3.3A), the BC-loop in almost every structure houses one residue, which points directly towards the heme. By a superposition of all structures, it could be shown that this position exactly is located at the same position, corresponding to the phenylalanine in P450 BM-3 (figure 6.3.3B) located at position 87. Table 6.3.2 lists the corresponding residue in each structure.

To validate our structure-based method to assign SCRs in a one-leave-out cross-validation, the position which corresponds to F87 in P450 BM-3 was predicted for each sequence of each structure. For 23 out of 30 (80 %), the predicted positions agreed with the crystal structure, in 7 CYPs they deviated by up to 2 residues. To further apply and to validate the procedure, the position was predicted in eleven structures published in progress of this study. 8 correct predictions, 2 deviations by one position, and one wrong prediction for the case of CYP7A1 which has in the crystal structure no residue located at this position, again confirmed an accuracy of 80 %. It should be noticed that the residue numbering of the structures of CYP2E1 and CYP74A does not start at 1 and therefore the numbering of the protein sequence was considered. The crystal structure of CYP231A thermoacidophilic *Picrophilus torridus* was missing a part of the BC-loop (Ho et al., 2008) which made the prediction not clearly defined.

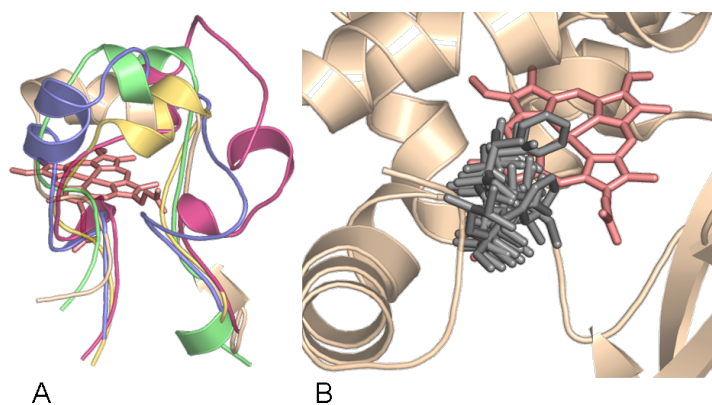


Figure 6.3.3: BC-loop region (SRS1) of CYPs. (A) Comparison of the BC-loops of P450 BM-3 (1BU7) in beige, CYP2C9 (1OG2) in green, CYP154C1 (1GWI) in pink, CYP101D (2CPP) in yellow and CYP107A1 (1OXA) in blue. (B) BC-loop region in P450 BM-3 (1BU7) and the position 87 corresponding residue in all 31 structures.

Table 6.3.2: Positions which correspond to F87 in P450 BM-3 and predicted position, and prediction of positions in new structures

CYP	PDB-Code	Position in crystal structure	Predicted position
8A	2IAG	-	-
51B1	1E9X	-	V88
55A2	1CL6	V87	V87
101D	2CPP	T101	I99
107A1	1OXA	G91	G91
107L1	2BVJ	L93	L93
108A	1CPT	T103	T103
119	1IO7	L69	L69
152A1	1IZO	Q85	Q85
154A1	1ODO	F88	F88
154C1	1GWI	L93	L93
158A1	2DKK	A97	S95
158A2	1S1F	G94	G94
165B3	1LFK	M89	N87
165C4	1UED	S98	S98
167A1	1Q5D	F96	G94
175A1	1N97	L80	L80
176A1	1T2B	A91	M89
199A2	2FR7	L100	L100
245A1	2Z3T	V99	V99
1A2	2HI4	T124	S126
2A6	1Z10	V117	V117
2A13	2P85	A117	A117
2B4	1SUO	I114	I114
2C5	1N6B	A113	A113
2C8	1PQ2	I113	I113
2C9	1OG2	V113	V113
2D6	2F9Q	F120	F120
2R1	2OJD	L125	L125
3A4	1TQN	S119	S119
102A1	1BU7 (reference)	F87	F87
2E1	3E4E	I94	I94
3A43	2V0M	S119	S119
7A1	2DAX	-	D98
19A1	3EQM	F134	F134
46A1	2Q9F	V126	S127
74A1	2RCH	S128	L127
105A1	2ZBX	I96	I96
105K1	2Z36	L96	L96
120A1	2VE3	A94	A94
231A2	2RFB	I48	I48
248A	3BUJ	L80	L80

Amino acid composition of the F87 corresponding position

In addition to the identification of the F87 corresponding position, a comprehensive analysis of the sequences of all 8614 *CYPED* protein entries was performed in respect to the amino acid composition, by a prediction of the position in all sequences analogous to the SCR prediction. It could be observed that 73 % of the residues predicted at this position include aliphatic residues and phenylalanine. The remaining 24 % at this position are small polar residues and only are 3 % charged residues. Phenylalanine (22 %), leucine (22 %), and valine (12 %) were the most frequently occurring amino acids followed by isoleucine (10 %) and alanine (9 %). Other amino acids appear more rarely with frequencies less than 4 % (figure 6.3.4).

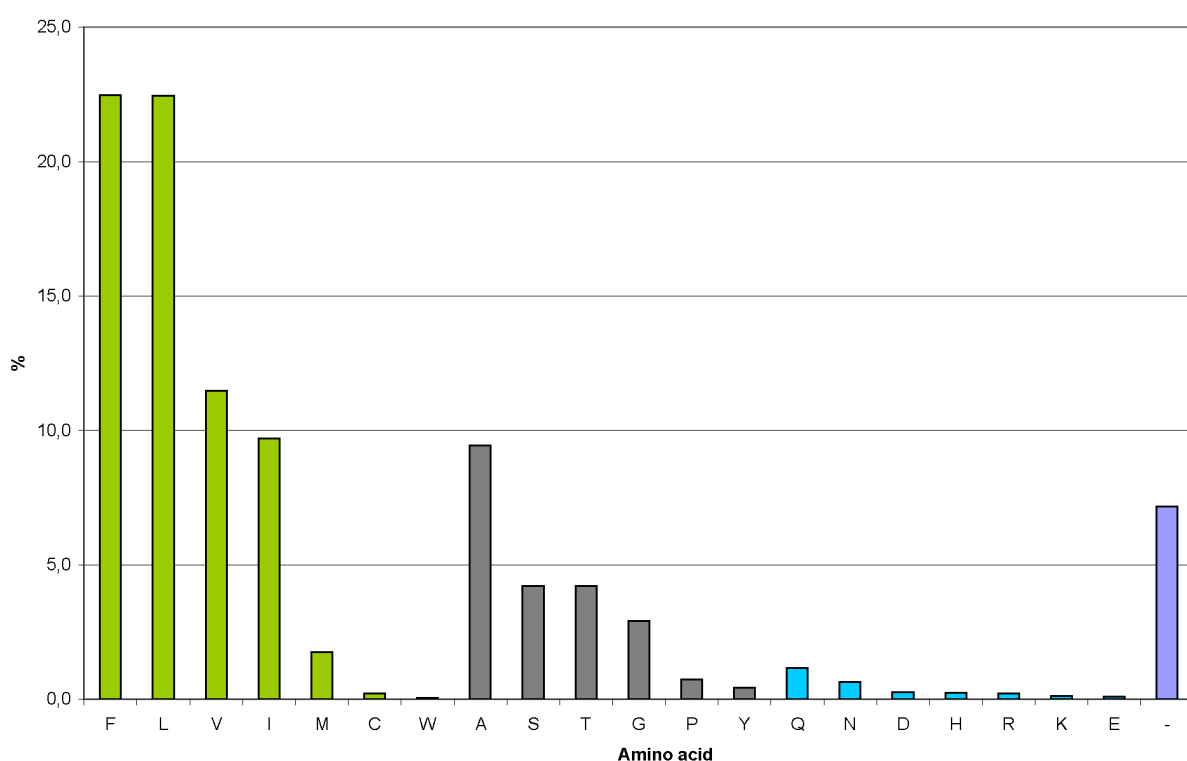


Figure 6.3.4: Amino acid composition of predicted F87 corresponding positions in all 8614 *CYPED* proteins. Green bars correspond to the percentage of aliphatic residues and phenylalanine, grey bars to amino acids of small polar nature and blue bars to charged amino acids. '-' denotes a gap in the alignment at the corresponding position.

Analysis of reductase interacting sites

The structural regions $\alpha J/J'$ and the insertion between the meander loop and the Cys-pocket are of particular interest since they were previously proposed to form the reductase interacting face of the molecules (Hasemann et al., 1995). These sites strongly vary

in their length and conformation. The structural analysis (figure 6.3.5) reflects the differences of $\alpha J/J'$ (further referred to as reductase interaction site 1, RIS1) (figure 6.3.5A) and the insertion between meander loop and Cys-pocket (further referred to as reductase interaction site 2, RIS2) (figure 6.3.5B) of CYPs from different redox classes. A comparison of the human CYP2C9 and the bacterial P450cam CYP101D shows that RIS1 ($\alpha J/J'$ region) of CYP2C9 is 18 residues longer. RIS2 differs by 9 residues between CYP2C9 and CYP101D. By counting the number of residues spanning these regions in the STAMP alignment (figure 6.3.7) was revealed that these regions in class I CYPs interacting with CPR-type reductases are long, in class II CYPs extremely short or not existing at all and in class III CYPs which do not require any electron transfer partner extremely long. The $\alpha J/J'$ region differs from 21 to 22 residues for class I and class III (long) and 3 to 5 residues for class II CYPs (short). The length of the meander insertion differs from 11 to 17 residues for class I (long), up to 23 residues for class III (very long) and 3 to 5 residues for class II CYPs (short).

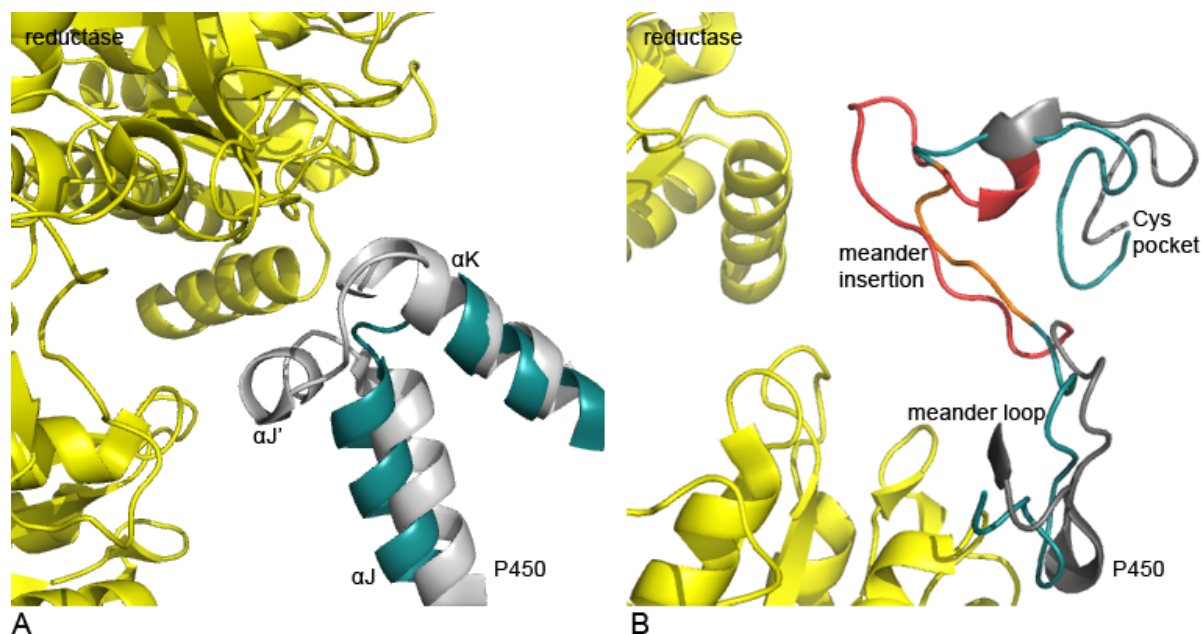


Figure 6.3.5: Sites interacting with potential redox partners. The CPR-type FMN/FAD (PDB: 3ES9 from *Rattus norvegicus*) is shown in yellow, the parts of the P450 domains are shown in grey (PDB: 1OG2 from *Homo sapiens*, CYP2C9) and green (PDB: 2CPP from *Pseudomonas putida*, CYP101D), respectively. (A) Comparison RIS1 ($\alpha J/J'$ region) of the human CYP2C9 and P450cam CYP101D. (B) Comparison of RIS2 (meander insertion) of the human CYP2C9 and P450cam CYP101D.

Counting the number of amino acids in each *CYPED* sequence for RIS1 (figure 6.3.6A) revealed two peaks in the RIR1 length distribution. This allowed to define two classes.

Proteins having short RIS1 with less than 10 residues spanning the $\alpha J/J'$ region make up 17.5 % of all protein entries. According to the result of the length analysis of RIS1 of the structural alignment, they comprise class II CYPs. Proteins having long RIS1 with more than 15 residues spanning the $\alpha J/J'$ region make up 81 % of all protein entries. According to the result of the length analysis of RIS1 of the structural alignment, they comprise class I and class III CYPs. Only 1 % of all protein entries can not reliably be assigned by RIS1 length since their length is in between 10 and 15 amino acids. 0.5 % of entries with RIS1 length above 35 amino acids were excluded from the analysis since they were considered as biologically not relevant.

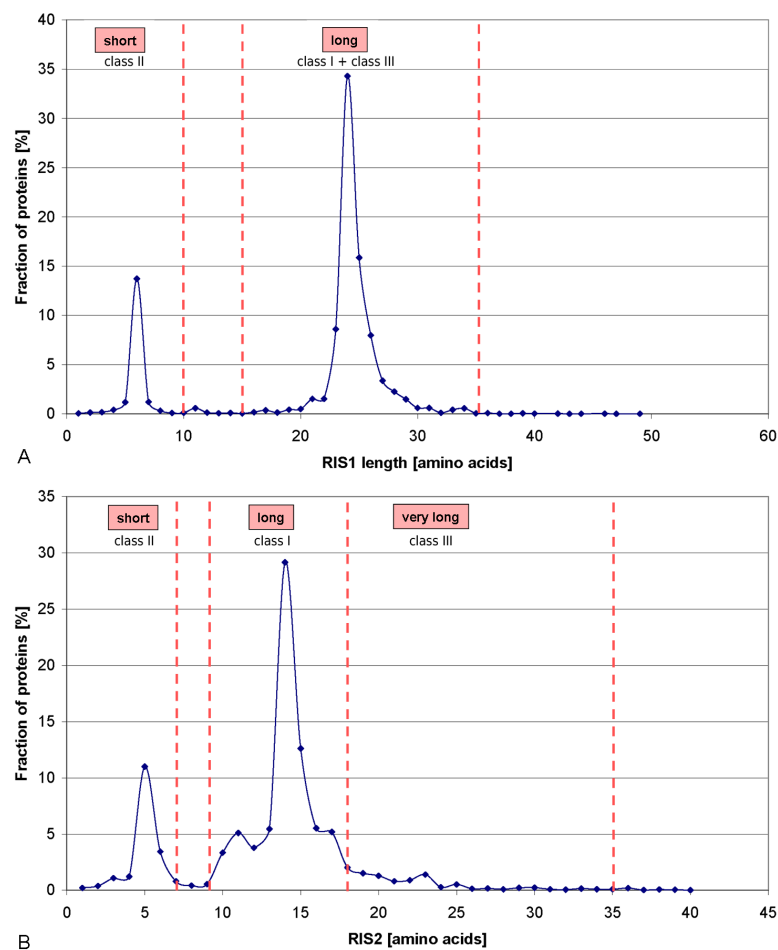


Figure 6.3.6: (A) Fraction of proteins for each RIS1 ($\alpha J/J'$ region) length. (B) Fraction of proteins for each RIS2 (meander insertion) length.

The analysis of the length of RIS2 in each *CYPED* sequence (figure 6.3.6B) showed a distribution in three main areas. Therefore, three classes according to the result of the length analysis of RIS2 in the structural alignment were defined. Proteins having

short RIS2 with less than 7 residues spanning the meander insertion make up 18 % of all protein entries. According to the result of the length analysis of RIS2 in the structural alignment, they comprise class II CYPs. Proteins having long RIS2 with between 11 and 17 residues spanning the meander insertion make up 66 % of all protein entries. According to the result of the length analysis of RIS2 in the structural alignment, they comprise class I CYPs. Proteins having very long RIS2 with more than 18 residues spanning the meander insertion make up 10 % of all proteins. According to the result of the length analysis of RIS2 in the structural alignment, they comprise class III CYPs. 4 % of all protein entries can not reliably be assigned by RIS2 length since their length is in between 8 and 10 amino acids. 0.5 % of entries with RIS2 length above 35 amino acids were excluded from the analysis since they were considered as biologically not relevant.

6.3.6 Discussion

Despite their inherently low sequence similarity, all CYPs share a common structural fold. The well-defined secondary structure elements can be found in all determined crystal structures, which house their active-site with the cofactor heme deeply inside the protein (Mestres, 2005). The generation of a structural alignment out of 31 CYP structures revealed structurally conserved regions which contain most of the described secondary structure elements of the CYP fold. It could be shown that some of the secondary structure elements merge together to structure modules, described as structurally conserved regions (SCR) 1-19, reflecting the modular structure of cytochrome P450 monooxygenases. The generation of a reliable structure-based HMM profile which was applied to every *CYPED* entry assisted in consistently annotating the conserved secondary structures in the *CYPED* entries. But besides addressing the problem of predicting conserved regions, an even more challenging issue could be solved: the identification and classification of the variable regions.

Since the residues that determine the substrate specificity of CYPs are assumed to lie in the variable regions (Peterson and Graham, 1998; Gotoh, 1992), their identification is of greatest interest for engineering of biochemical properties. Two of the six proposed substrate recognition sites, SRS1 and SRS5, together with the helix I directly flank the substrate binding cavity and are therefore supposed to interact with the substrate (Seifert and Pleiss, 2008). SRS1 houses a residue, which previously was described as essential for activity, regio- and stereoselectivity in P450 BM-3 (Li et al., 2008; Urlacher and Schmid, 2002; Urlacher et al., 2006). Located at position 87 and pointing directly towards the

heme, a corresponding residue to this phenylalanine can be found in almost all CYP structures. Its location in the highly variable BC-loop region makes its determination very difficult in sequences without structural information.

The position, which corresponds to F87 in P450 BM-3 could be correctly predicted in almost 80 % of all analyzed CYP structures. By surveying more recent CYP structures, the validity of the prediction could be confirmed. The analysis of this position in all 8614 CYP sequences in the *CYPED* revealed that the residues at this position predominantly are of aliphatic nature or a phenylalanine, less frequently small polar amino acids and only very infrequently of charged nature. Since the characteristics of the residue at this position highly influence substrate specificity and regioselectivity, its identification contributes to the design of CYPs with more suitable properties for biocatalytic applications.

Even though there were two reductase interaction sites proposed to be located in $\alpha J/\alpha J'$ and in the insertion following the meander loop (Hasemann et al., 1995), termed RIS1 and RIS2, these regions which are highly variable in sequence and structure were difficult to determine in sequences. The identification of the preceding and the successive SCR solved this problem. Depending on the length for RIS1, two classes (short and long RIS1) and for RIS2 three classes (short, long and very long RIS2) were introduced. From the analysis of the CYP structures in respect to their redox partner it was assumed that class I CYPs have long RIS1 and long RIS2, class II CYPs have short RIS1 and short RIS 2, and that CYPs belonging to class III have long RIS1 and very long RIS2.

The largest percentage of all CYPs has long RIS1 and long RIS2 (53 %). All CYPs with available structure which possess these long loops clearly belong to class I, and most of them are of human origin. The class I protein P450 BM-3 also shows the characteristic CPR-interacting loop length. The 12 % of proteins with short RIS1 and RIS2, respectively, are assumed to be class II proteins. CYPs with long RIS1 and very long RIS2 make up the smallest percentage of all CYPs (8 %). The remaining 27 % could not be clearly classified, either because of unusual long loops (above 35 residues), or combine short RIS1 with long RIS2 and vice versa. This comparison of reductase interaction site allows to draw conclusion on its reductase interaction. The human prostacyclin synthase CYP8A1, which has endoperoxidase activity and does not require a reductase as source of electrons, is a representative of this class (Chiang et al., 2006). It has a long

RIS1, consisting of 22 amino acids and a very long RIS2 of 23 amino acids. The crystal structure for the human cholesterol 7 alpha-hydroxylase CYP7A1 which was recently solved also contains very long proximal loops (Strushkevich et al., PDB: 2DAX) which were correctly predicted containing 22 (RIS1) and 23 (RIS2) amino acids. CYP7A1 was previously compared to the structure of CYP8A1 (Mast et al., 2005). The fatty acid hydroxylase CYP152A1 from *Bacillus subtilis* (P450_{Bsβ}) is a hydrogen peroxide driven enzyme (Lee et al., 2003) and therefore belongs to class III CYPs. It has a short RIS1 of 5 amino acid residues and a long, and not as expected for the class III CYPs, RIS2 of 11 residues, like the CPR-type interacting class II CYPs. Indeed, this enzyme and its homologous protein CYP152A2 from *Clostridium acetobutylicum* (P450_{CLA}) experimentally showed much higher conversions in the presence of a CPR-type reductase than in the presence of hydrogen peroxide and the absence of a reductase (Girhard et al., 2007).

Since most CYPs require electrons from a redox partner, and even those who do not in some cases showed higher activities by adding a reductase, it can be assumed that the interaction of CYPs with reductases plays a pivotal role in the CYP mechanism. Finding the optimal redox partner for CYPs may significantly enhance their activity but is quite difficult. The analysis and classification which led to the prediction of possible redox partner interactions offers the potential of engineering enhanced interactions.

6.3.7 Conclusion

In order to navigate in all CYP sequences and to determine functionally relevant residues, a procedure which allows identifying conserved modules and functionally relevant sites within variable regions was implemented. Regions involved in substrate binding as well as redox partner recognition and interaction could be determined in the absence of structural information, based on sequence only. The structurally annotated sequences and multisequence alignments are accessible on the current version of the *CYPED* <http://www.cyped.uni-stuttgart.de>. Via a web interface integrated in the *CYPED* homepage at <http://www.cyped.uni-stuttgart.de/cgi-bin/strpred/dosecpred.pl>, the structural prediction is provided for every sequence which is similar to CYPs or presumably shares the CYP fold. The navigation in CYP sequences and the determination of functionally relevant sites in turn is a great advantage in the prediction of promising targets for the design of CYPs with improved biocatalytic properties.

6.3.8 Authors' contributions

DS implemented the program, performed the analysis and wrote the manuscript. FW carried out the annotation and generated the web interface. MWI contributed to the analysis and to the manuscript. JP supervised the project and finalized the manuscript.

6.3.9 Acknowledgements

This work was financially supported by the Deutsche Forschungsgemeinschaft (SFB 706).

6.3.10 Supplementary material

2iaga	1	R	10	G	20	L	30	G	40	D	F	50	G	K	60	70
1e9xa	MS	T	R	P	A	V	A	L	D	E	L	F	R	T	D	A
1z0a																
1cpta	M			D	H	G	A									
2i7a																
1i07a																
1i1ka																
1ueda	D		I	D	V	A	P		L							
1c16a																
1s11a																
2dkka				Q	A											
2z3ta																
1gwia																
1q5da																
1od0a																
1oxaa																
2bvja																
1i2ba																
2cppa	N	L	A	P	L	P	P	H	V	P	P					
2h4a	R	V	P	K	G	L	K	S								
2f9qa																
2ojda																
1su0a																
1z10a																
2p85a	K															
1n6ba																
1og2a																
1pc2a																
1tqna	HS															
1bu7a																
1n97a																

SCR1			SCR2			SCR3			SCR4			
αA			β1-1			β1-2	αB		β1-5			
2iaga	A	S	F	L	T	R	R	A	E	F	I	D
1e9xa	I	G	L	M	Q	R	R	A	F	F	R	R
1z0a	Y	F	I	K	N	R	T	E	E	Y	N	
1cpta	Y	P	A	F	K	W	L	R	E	E	A	
2i7a	Y	P	E	Q	E	S	E	L	R	E	A	
1i07a	Y	D	W	F	L	L	A	M	R	K	K	
1i1ka	A	D	E	L	L	L	A	K	G			
1ueda	H	E	D	N	F	A	K	G	L	R	A	H
1c16a	P	A	E	F	L	T	E	E	Y	N		
1s11a	D	P	V	L	A	E	L	M	R	E	E	
2dkka	S	K	P	V	L	A	E	L	E	R	A	
2z3ta	Y	P	P	V	Y	R	R	Y	R	E	A	
1gwia	A	R	L	R	A	A	L	R	E	A		
1q5da	F	P	A	I	E	R	E	S				
1od0a	H	R	T	L	R	E	E	G				
1oxaa	F	E	A	L	R	A	E	T				
2bvja	Y	F	E	T	Y	A	R	O				
1i2ba	W	A	V	L	A	E	S					
2cppa	F	E	A	L	O	E	S					
2h4a	H	L	A	L	S	R	O	R	Y			
2f9qa	P	Y	C	F	D	O						
2ojda	H	V	Y	M	R	K	R	O	S	O	Q	
1su0a	L	R	S	F	L	R						
1z10a	Y	N	S	L	M	K	I	S	E	R	R	
2p85a	Y	N	S	L	M	K	I	S	E	R	R	
1n6ba	S	K	S	L	T	K						
1og2a	S	K	S	L	T	N						
1pc2a	C	K	S	F	T	N						
1tqna	C	M	F	D	M	E						
1bu7a	V	O	A	L	M	K						
1n97a	L	A	V	L	L	A						

2iaga	150	160	170	180	190	200	210
1e9xa				Y	A	I	F
1z0a				L	E	I	L
1cpta	F	K					
2i7a	T	G	Y	H	E	R	L
1i07a							
1i1ka							
1ueda	T	R					
1c16a	M	D					
1s11a	T	Q					
2dkka							
2z3ta							
1gwia	N	V					
1q5da	E	E					
1od0a	R	A	H				
1oxaa	K	K					
2bvja	R	N					
1i2ba							
2cppa							
2h4a							
2f9qa							
2ojda							
1su0a							
1z10a							
2p85a							
1n6ba							
1og2a							
1pc2a							
1tqna							
1bu7a							
1n97a							

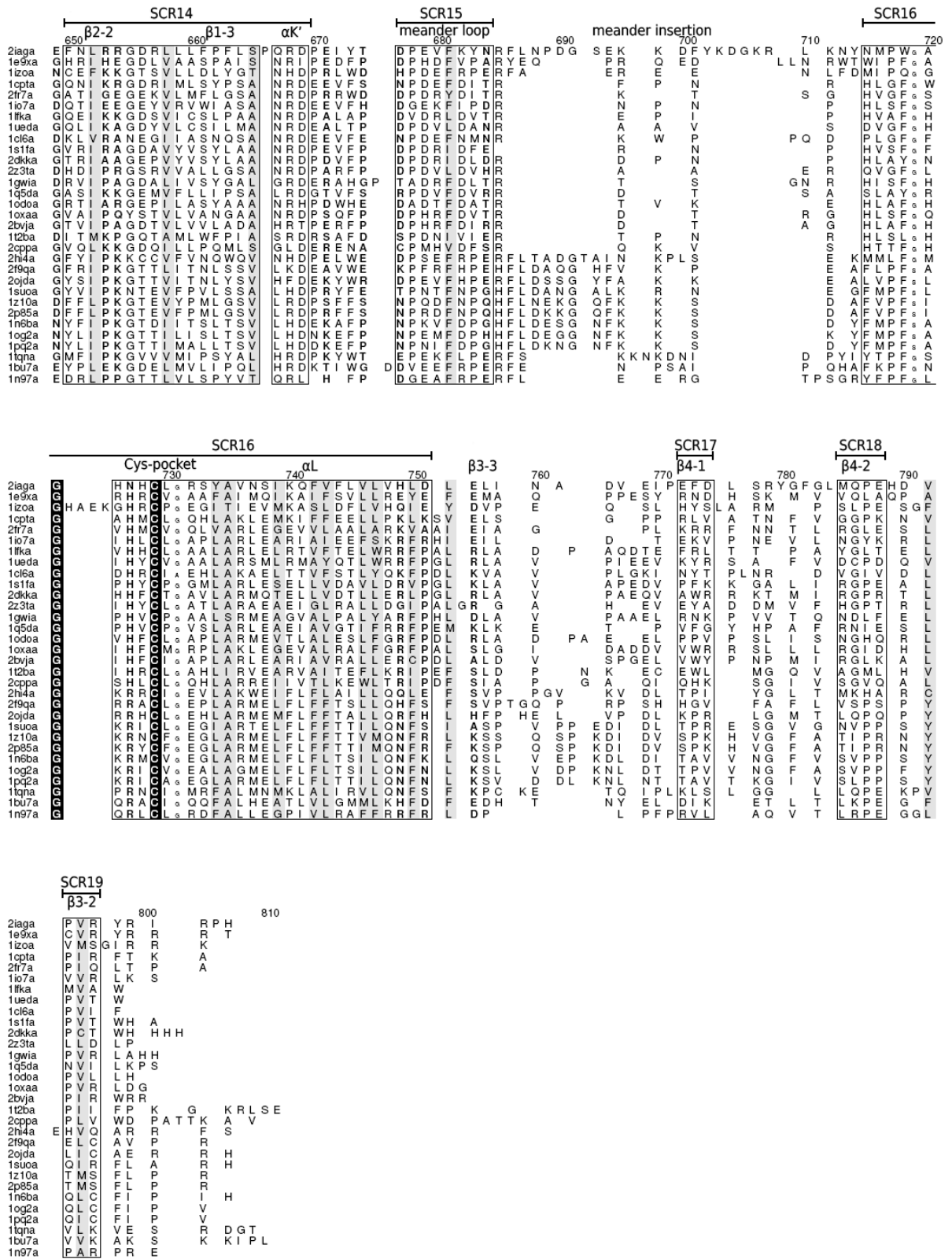


Figure 6.3.7: STAMP alignment output. Stretches of residues with STAMP score $S_c > 6.0$ imply regions of conserved functions and are marked in black boxes.

6.3.11 References

- Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25 (17), 3389–3402, 1997.
- Baudry, J., Rupasinghe, S., Schuler, M. A., Class-dependent sequence alignment strategy improves the structural and functional modeling of P450s. *Protein Eng Des Sel* 19 (8), 345–353, 2006.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L., GenBank. *Nucleic Acids Res* 36 (Database issue), D25–D30, 2008.
- Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., Zardecki, C., The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 58 (Pt 6 No 1), 899–907, 2002.
- Bernhardt, R., Cytochrome P450: structure, function, and generation of reactive oxygen species. *Rev Physiol Biochem Pharmacol* 127, 137–221, 1996.
- Bernhardt, R., Cytochromes P450 as versatile biocatalysts. *J Biotechnol* 124 (1), 128–145, 2006.
- Chiang, C., Yeh, H., Wang, L., Chan, N., Crystal Structure of the Human Prostacyclin Synthase. *J. Mol. Biol.* 364, 266–274, 2006.
- de Graaf, C., Vermeulen, N. P. E., Feenstra, K. A., Cytochrome p450 in silico: an integrative modeling approach. *J Med Chem* 48 (8), 2725–2755, 2005.
- Delano, W. L., The PyMOL Molecular Graphics System. San Carlos, CA, USA: DeLano Scientific, 2002.
- Fischer, M., Knoll, M., Sirim, D., Wagner, F., Funke, S., Pleiss, J., The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. *Bioinformatics* 23 (15), 2015–2017, 2007.
- Fischer, M., Thai, Q. K., Grieb, M., Pleiss, J., DWARF—a data warehouse system for analyzing protein families. *BMC Bioinformatics* 7, 495, 2006.

- Girhard, M., Schuster, S., Dietrich, M., Dürre, P., Urlacher, V. B., Cytochrome P450 monooxygenase from *Clostridium acetobutylicum*: a new alpha-fatty acid hydroxylase. *Biochem Biophys Res Commun* 362 (1), 114–119, 2007.
- Gotoh, O., Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J Biol Chem* 267 (1), 83–90, 1992.
- Graham, S. E., Peterson, J. A., How similar are P450s and what can their differences teach us? *Arch Biochem Biophys* 369 (1), 24–29, 1999.
- Guengerich, F. P., Rate-limiting steps in cytochrome P450 catalysis. *Biol Chem* 383 (10), 1553–1564, 2002.
- Hannemann, F., Bichet, A., Ewen, K. M., Bernhardt, R., Cytochrome P450 systems—biological variations of electron transport chains. *Biochim Biophys Acta* 1770 (3), 330–344, 2007.
- Hasemann, C. A., Kurumbail, R. G., Boddupalli, S. S., Peterson, J. A., Deisenhofer, J., Structure and function of cytochromes P450: a comparative analysis of three crystal structures. *Structure* 3 (1), 41–62, 1995.
- Ho, W. W., Li, H., Nishida, C. R., de Montellano, P. R. O., Poulos, T. L., Crystal structure and properties of CYP231A2 from the thermoacidophilic archaeon *Picrophilus torridus*. *Biochemistry* 47 (7), 2071–2079, 2008.
- Kabsch, W., Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12), 2577–2637, 1983.
- Lee, D.-S., Yamada, A., Sugimoto, H., Matsunaga, I., Ogura, H., Ichihara, K., Adachi, S.-I., Park, S.-Y., Shiro, Y., Substrate recognition and molecular mechanism of fatty acid hydroxylation by cytochrome P450 from *Bacillus subtilis*. Crystallographic, spectroscopic, and mutational studies. *J Biol Chem* 278 (11), 9761–9767, 2003.
- Li, H. M., Mei, L. H., Urlacher, V. B., Schmid, R. D., Cytochrome P450 BM-3 evolved by random and saturation mutagenesis as an effective indole-hydroxylating catalyst. *Appl Biochem Biotechnol* 144 (1), 27–36, 2008.
- Mast, N., Graham, S. E., Andersson, U., Bjorkhem, I., Hill, C., Peterson, J., Pikuleva, I. A., Cholesterol binding to cytochrome P450 7A1, a key enzyme in bile acid biosynthesis. *Biochemistry* 44 (9), 3259–3271, 2005.

- McLean, K. J., Sabri, M., Marshall, K. R., Lawson, R. J., Lewis, D. G., Clift, D., Balding, P. R., Dunford, A. J., Warman, A. J., McVey, J. P., Quinn, A. M., Sutcliffe, M. J., Scrutton, N. S., Munro, A. W., Biodiversity of cytochrome P450 redox systems. *Biochem Soc Trans* 33 (Pt 4), 796–801, 2005.
- Mestres, J., Structure conservation in cytochromes P450. *Proteins* 58 (3), 596–609, 2005.
- Montellano, O. d., *Cytochrome P450: structure, mechanism and biochemistry*. New York, Plenum Press, 1995.
- Munro, A. W., Girvan, H. M., McLean, K. J., Cytochrome P450-redox partner fusion enzymes. *Biochim Biophys Acta* 1770 (3), 345–359, 2007.
- Munro, A. W., Leys, D. G., McLean, K. J., Marshall, K. R., Ost, T. W. B., Daff, S., Miles, C. S., Chapman, S. K., Lysek, D. A., Moser, C. C., Page, C. C., Dutton, P. L., P450 BM3: the very model of a modern flavocytochrome. *Trends Biochem Sci* 27 (5), 250–257, 2002.
- Nelson, D. R., Cytochrome P450 nomenclature, 2004. *Methods Mol Biol.* 320, 1–10, 2006.
- Peterson, J. A., Graham, S. E., A close family resemblance: the importance of structure in understanding cytochromes P450. *Structure* 6 (9), 1079–1085, 1998.
- Picard, R. R., Cook, R. D., Cross-Validation of Regression Models. *Journal of the American Statistical Association* 79 (387), 575–583, 1984.
- Raucy, J. L., Allen, S. W., Recent advances in P450 research. *Pharmacogenomics J* 1 (3), 178–186, 2001.
- Rossmann, M. G., Argos, P., Exploring structural homology of proteins. *J Mol Biol* 105 (1), 75–95, 1976.
- Russell, R. B., Barton, G. J., Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* 14 (2), 309–323, 1992.
- Seifert, A., Pleiss, J., Identification of selectivity-determining residues in cytochrome P450 monooxygenases: A systematic analysis of the substrate recognition site 5. *Proteins*, 2008.

- Seifert, A., Vomund, S., Grohmann, K., Kriening, S., Urlacher, V. B., Laschat, S., Pleiss, J., Rational design of a minimal and highly enriched CYP102A1 mutant library with improved regio-, stereo- and chemoselectivity. *Chembiochem* 10 (5), 853–861, 2009.
- Sirim, D., Wagner, F., Lisitsa, A., Pleiss, J., The cytochrome P450 engineering database: Integration of biochemical properties. *BMC Biochem* 10, 27, 2009.
- Smith, T. F., Waterman, M. S., Identification of common molecular subsequences. *J Mol Biol* 147 (1), 195–197, 1981.
- Strushkevich, N., Tempel, W., Dombrovski, L., Dong, A., Loppnau, P., Arrowsmith, C., Edwards, A., Bountra, C., Wilkstrom, M., Bochkarev, A., Park, H., Crystal structure of human CYP7A1. To be Published, PDB: 2DAX.
- Urlacher, V., Schmid, R. D., Biotransformations using prokaryotic P450 monooxygenases. *Curr Opin Biotechnol* 13 (6), 557–564, 2002.
- Urlacher, V. B., Eiben, S., Cytochrome P450 monooxygenases: perspectives for synthetic application. *Trends Biotechnol* 24, 324–330, 2006.
- Urlacher, V. B., Makhsumkhanov, A., Schmid, R. D., Biotransformation of beta-ionone by engineered cytochrome P450 BM-3. *Appl Microbiol Biotechnol* 70 (1), 53–59, 2006.
- Wade, R. C., Motiejunas, D., Schleinkofer, K., Sudarko, Winn, P. J., Banerjee, A., Kariakin, A., Jung, C., Multiple molecular recognition mechanisms. Cytochrome P450—a case study. *Biochim Biophys Acta* 1754 (1-2), 239–244, 2005.
- Werck-Reichhart, D., Feyereisen, R., Cytochromes P450: a success story. *Genome Biol* 1 (6), REVIEWS3003, 2000.

6.4 Immobilization of P450 BM-3 monooxygenase on mesoporous molecular sieves with different pore diameters

Erschienen in *Journal of Molecular Catalysis B - Enzymatic* **64**: 29-37.

Weber, E., Sirim, D., Schreiber, T., Thomas, B., Pleiss, J., Hunger, M., Gläser, R., Urlacher, V.B., 2010. Immobilization of P450 BM-3 on mesoporous molecular sieves.

Immobilization of P450 BM-3 monooxygenase on mesoporous molecular sieves with different pore diameters

Evelyn Weber^a, Demet Sirim^a, Tino Schreiber^{b,1}, Bejoy Thomas^b, Jürgen Pleiss^a, Michael Hunger^b, Roger Gläser^c, Vlada B. Urlacher^{a,*}

^a Institute of Technical Biochemistry, Universitaet of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany
^b Institute of Chemical Technology, Universitaet of Stuttgart, Pfaffenwaldring 55, 70569 Stuttgart, Germany
^c Institute of Chemical Technology, University of Leipzig, Linnéstraße 3, 04103 Leipzig, Germany

ARTICLE INFO

Article history:
 Received 9 October 2009
 Received in revised form 22 January 2010
 Accepted 22 January 2010
 Available online 1 February 2010

Keywords:
 P450 BM-3
 Mesoporous molecular sieves
 SBA-15
 MCM-41
 Immobilization

ABSTRACT

The immobilization of the isolated heme domain of P450 BM-3 (BM3H.F87A) on two mesoporous molecular sieves, MCM-41 (pore diameter 25 Å) and SBA-15 (pore diameter 60 Å and 133 Å) was examined systematically, and the activity of the immobilized enzyme toward *para*-nitrophenoxycarboxylic acid (12-*p*NCA) and *n*-octane was determined. Hydrogen peroxide was utilized as source of electrons and oxygen to support the monooxygenase activity of BM3H.F87A. The mesoporous materials were characterized by X-ray diffraction and nitrogen adsorption analyses before and after immobilization. The results revealed that the immobilization efficiency of MCM-41 and SBA-15 after single immersion was strongly affected by the pH value of the enzyme solution, initial enzyme concentration and agitation conditions. By modelling the 3D structure *in silico* and performing electrostatic potential calculations, the pH-dependence of the enzyme immobilization could be explained and a possible orientation of the protein on mesoporous materials was predicted. The oxidizing activity of the immobilized enzyme was found to depend on pore diameter and accessibility of the substrate for the enzyme. The highest activity toward 12-*p*NCA of 830 nmol product/mg P450/min was observed with BM3H.F87A immobilized on SBA-15 with pore diameter 133 Å. Enzyme activity toward *n*-octane was similar for the enzyme immobilized on SBA-15 of 60 Å and 133 Å, and was at least two-fold higher as compared to a system with free enzyme.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

In general, immobilization with respect to enzyme stabilization refers to associating an enzyme with an insoluble matrix, so that it can be reused under stabilized conditions. Immobilized biocatalysts offer several other advantages, like improved enzyme storage and operational stability, resistance to elevated temperatures and organic (co)-solvents, the possibility for continuous processes and greater control over enzymatic reactions. Although immobilization to solid carriers is perhaps the most frequently used strategy to improve the stability of enzymes, only a few reports of successful cytochrome P450 immobilization can be found in literature. Cytochrome P450 enzymes (P450s) are heme containing proteins that catalyze oxygenation of a vast variety of organic molecules. One of the problems regarding the use and immobiliza-

tion of P450 enzymes is their dependency on the pyridine co-factors NADH and NADPH and the need for the corresponding reductases, which transfer electrons from NAD(P)H to the heme group. The first example on immobilization of P450s dates back to 1988 when Wiseman and co-workers [1] immobilized purified P450s from *Saccharomyces cerevisiae* along with the corresponding reductase by entrapment in calcium alginate or in polyacrylamide, or by adsorption on cyanogen bromide-activated sepharose. A decade later, the plant CYP71B1, fused to a P450 reductase, was immobilized onto colloidal liquid aphrons [2]. Kelly and co-workers reported the co-immobilization of prokaryotic CYP105D1 with a ferredoxin onto the ionic exchange resin DE52-72 [3]. However, most of these systems suffered from the leaching of the enzyme activity from the support.

Previously we have reported the immobilization of the P450 BM-3 from *Bacillus megaterium* (CYP102A1) and some of its mutants on different supports [4]. P450 BM-3 monooxygenase is a self-sufficient natural fusion flavocytochrome (119 kDa) consisting of a heme domain and a diflavin reductase domain [5]. The enzyme was unable to adsorb neither on celite, a porous silicate matrix derived from diatomaceous soil, nor on Eupergit C via covalent

* Corresponding author.

E-mail address: Vlada.Urlacher@itb.uni-stuttgart.de (V.B. Urlacher).

¹ Institute for Interfacial Engineering, Universitaet of Stuttgart, Nobelstr. 12, 70569 Stuttgart, Germany.

binding. Negative results obtained with polypropylene derivatives and phenyl-, octyl- and butylsepharose showed that procedures based on hydrophobic interactions were also unsuitable for the efficient immobilization of P450 BM-3. P450 BM-3 binds to anion exchangers such as DEAE and SuperQ [4]. However, the most substrates and products of the oxidation reaction also adsorbed on these matrices. In addition, enzyme leaching from the carrier occurs in buffers with high ionic strength. Furthermore, the purified P450 BM-3 A74G/F87V/L188Q mutant was successfully encapsulated in a sol-gel matrix derived from tetraethoxyorthosilicate (TEOS) upon polymerization. The entrapment of P450 BM-3 in this sol-gel-type material resulted in a very high long-term storage stability of the enzyme at different temperatures. A half-life of 29 days was measured at 25 °C for immobilized P450 BM-3 and only 2 days for the free enzyme. A sol-gel immobilized P450 BM-3 mutant was able to oxidize substrates of diverse substance classes such as terpenoids, polyaromatic hydrocarbons, *n*-alkanes, and fatty acid analogs with high activity [4]. However, since the entrapment of P450 BM-3 was performed during polymerization of TEOS, we were not able to control pore diameter and pore geometry properly. The use of ordered mesoporous silicates, synthesized using surfactant templating routes and therefore having defined pore diameter and pore geometry, would clarify these aspects. Since the first report by Diaz and Balkus in 1996 [6] various commercially available enzymes such as cytochrome *c* [7], lysozyme [8], lipase [9,10] and albumin [11] have successfully been immobilized on ordered mesoporous materials like MCM-41, MCM-48 (Mobil Composition of Matter) SBA-15 or SBA-16 (Santa Barbara). In some cases proteins were adsorbed only on the external surface area of mesoporous materials, in the others – immobilized within their pores. In the meantime only two reports dedicated to P450 monooxygenases were published [12,13]. Both reports describe immobilization of rabbit CYP2C9 and human CYP2B4 on aluminum-substituted MCM-41 containing aluminum ions at different ratios. Interestingly, catalytic activity of immobilized CYP2C9 and CYP2B4 was observed even in the absence of the cytochrome P450 reductase, which is necessary for electron transfer from NADPH to the heme group. The authors suggested that electron transfer to the immobilized P450s can occur through the Lewis acid, i.e., the Al-centers in the silicate walls [12].

In the present study the immobilization of the isolated heme domain of P450 BM-3 (without the reductase domain) on two mesoporous ordered materials, MCM-41 (pore diameter 25 Å) and SBA-15 (pore diameter 60 Å and 133 Å), was systematically examined. For comparison, a commercial silica gel with a broad pore diameter distribution was included into this study, too. The aim of this study was to develop an effective immobilization procedure and to investigate the effect of the pore diameter on the loading capacity of both materials and oxygenase activity of the immobilized P450 BM-3 heme domain.

2. Experimental

2.1. Materials

All chemicals reagents were of analytical grade purity and purchased from Roth (Karlsruhe, Germany) and Fluka (Steinheim, Germany). H₂O₂ was purchased as a 30 wt.% solution from Fluka. The stock solution was prepared freshly in 50 mM potassium phosphate buffer, pH 7.5. Silica gel-Type 62 and sodium water glass (25.5–28.5 wt.% SiO₂, 7.5–8.5 wt.% Na₂O, rest: water) were obtained from Merck (Germany). 12-*para*-nitrophenoxydodecanoic acid (12-*p*NCA) was synthesized as described elsewhere [14] and dissolved in dimethyl sulfoxide (DMSO).

2.2. Preparation and mutagenesis of the P450 BM-3 heme domain

In our previous work the gene CYP102A1 encoding the cytochrome P450 BM-3 has been amplified from genomic DNA of *B. megaterium* ATCC 14581 and cloned into the pET28a(+)-vector yielding the pET-28a.BM-3 construct [15]. The gene fragment coding for the P450 BM-3 heme domain was amplified from pET-28a.BM-3 using the following primers: 5'-3': CCGGATCCATGACAATTAAGAAATGCCTCAGC; 5'-3': GCGAATTCTTAGCGTACTTTTTAGCAGACTGTC. The primer for the 3'-end of the gene contains an additional stop codon. The amplified gene as well as the pET28a(+)-vector were cut using the endonucleases *Bam*HI and *Eco*RI and then ligated together by T4-DNA ligase. The replacement of the phenylalanine at position 87 by smaller alanine was performed using the Quick-Change Kit (Stratagene). The primers were as follows: 5'-3': gcaggagacgggttagctcacagctggagc and 5'-3': gcgtccagctttagctaacctctctctgc. PCR was carried out following the manufacturer's protocol. The correct gene insertion and mutation were checked by sequencing. The His6-tagged P450 BM-3 F87A heme domain (further referred to as BM3H.F87A) was expressed in *Escherichia coli* BL21(DE3) and purified on Ni-NTA sepharose as described previously for the holoprotein [4]. The P450 concentrations were quantified from the CO-binding difference spectra of the reduced form as described elsewhere [16]. The extinction coefficient of 91 mM⁻¹ cm⁻¹ was used for calculations.

2.3. Immobilization of the P450 F87A heme domain on mesoporous materials

If not stated otherwise 1.5 mL of purified BM3H.F87A with a final concentration of 15–150 μM was added to 20 mg of a mesoporous material. The immobilization procedure was optimized upon different agitation conditions such as stirring, slow rotation at 15 rpm or intensive mixing. Experiments with stirring were carried out with a magnetic stirrer in a covered beaker. Experiments with rotation were carried out in a Rotamix (RM-1, ELMI, Latvia). Intensive mixing was performed in a Beadmill (MM2000, Retsch, Haan, Germany). Immobilization was performed during 1–24 h at 10 °C. The solids with the immobilized enzyme were recovered by centrifugation (10 min, 2000 g, 4 °C). The supernatant was used for estimation of the concentration of non-immobilized active P450 by measuring CO-difference spectra. The recovered solid fractions were washed four times with 50 mM potassium phosphate buffer, pH 7.5, and were stored then at –20 °C before use. CO-difference spectra measured with washing solutions were used for estimation of enzyme leaching. For sequential immersion the already one or more time loaded portion of a mesoporous material was reloaded again under equal conditions. The recovered solid was washed four times with 50 mM potassium phosphate buffer, pH 7.5 between reloading steps.

2.4. Activity toward *p*-nitrophenoxydodecanoic acid (12-*p*NCA)

P450 BM-3 activity assays were performed using the *p*NCA assay [14]. The reaction was carried out at room temperature in a final volume of 1.0 mL containing 50 mM potassium phosphate buffer, pH 8.1, 200 μM 12-*p*NCA dissolved in DMSO (final concentration 1%), and the corresponding amount of the purified enzyme. The reaction was started by adding 10 mM H₂O₂. Formation of *p*-nitrophenolate was followed at 410 nm on an Ultraspec 3000 photometer (Pharmacia Biotech, Uppsala, Sweden) and calculated using extinction coefficient of 13.2 mM⁻¹ cm⁻¹. Activity measurements with immobilized BM3H.F87A (20 mg for each experiment) were carried out under stirring in a flow-through cuvette at room temperature. Formation of *p*-nitrophenolate was followed with a Nicolet evolution 1000 photometer (Thermo Electron Corpo-

ration). All activity measurements were carried out at least in triplicate.

For the identification of enzyme leaching under process conditions the reaction mixture (without the loaded materials) was taken from the flow-through cuvette after 3 and 5 min of the reaction, and used for measuring CO-difference spectra and activity of the enzyme toward lauric acid. The pNCA test could not be applied here because the solution was already yellow. The reaction with lauric acid was monitored by GC–MS as described elsewhere [17].

2.5. Conversion of *n*-octane with immobilized P450 F87A heme domain

To maintain the equal P450 concentration in all experiments, the respective amount of immobilized BM3H.F87A was suspended in 1 mL 50 mM KPi pH 7.5, supplemented with 20 μ L of a 10 mM *n*-octane solution in ethanol and 10 mM H₂O₂. After 2 h the reaction mixture was centrifuged, and the supernatant was extracted twice with 300 μ L dichloromethane. The combined organic layers were supplemented with internal standard 1-decanol, dried over magnesium sulphate and concentrated to a volume of 100 μ L. Reaction products and unreacted substrate were measured on GC/MS (Shimadzu, Japan) using a FS-Supreme-5 column and identified by MS. The temperature gradients were as follows: (1) 40 °C for 1 min, (2) 40–67 °C at 2 °C/min, (3) 67–75 °C at 1 °C/min, (4) 75–280 °C at 30 °C/min. Pure samples of the substrate and potential reaction products 2-, 3-, and 4-octanols were available. Equal amounts of these substances dissolved in dichloromethane were applied to the column. From the resulting GC/MS trace the ratio of the peak areas corresponding to the substrates and products were calculated. These ratios were used to determine the molar ratios of substrates and products emerging from the biotransformations. Therefore equal dichloromethane–water partition coefficients for educts and products were assumed.

2.6. Synthesis of ordered mesoporous materials

2.6.1. MCM-41

MCM-41 was synthesized according to a modified procedure reported in literature [18]. Briefly, 8.13 g sodium water glass was added to 120 g demineralized water and stirred for 30 min. A second solution was prepared by dissolving 4.48 g tetradecyltrimethylammonium bromide in 30 g demineralized water, addition of 10 g ethanol and stirring for 30 min. The two solutions were combined under stirring to obtain a clear gel. After additional stirring for 30 min, 15.0 g 4N H₂SO₄ were added slowly and stirring was continued for 1 h. The resulting gel was transferred into Teflon-lined stainless steel autoclaves and kept at 150 °C for 20 h. After cooling, the solid product was separated by filtration, washed thoroughly with hot water (80 °C) and ethanol and, then, calcined at 550 °C first in a nitrogen atmosphere for 12 h and, subsequently, in air for 6 h.

2.6.2. SBA-15 (60 Å)

For the synthesis of SBA-15, a procedure reported by Choi et al. [19] was followed. Accordingly, 6.92 g of the tri-block copolymer (poly(ethylene oxide)-poly(propylene oxide)-poly(ethylene oxide)) P123 was dissolved in 43.2 g demineralized water and 8.75 g concentrated aqueous hydrochloric acid. Another solution was prepared by diluting 25.9 g sodium water glass with 69.5 g demineralized water and dissolving 0.27 g NaOH. This solution was added to the acidic tri-block copolymer solution at 35 °C under stirring. Stirring was continued for 24 h and, thereafter, the gel was placed in a Teflon-lined autoclave which was heated for 24 h at 100 °C. The solid product was separated, washed and calcined as described for MCM-41 above.

2.6.3. SBA-15 (133 Å)

Large pore SBA-15 was synthesized according to Vinu and co-workers [20]. Typically, 4 g of the tri-block copolymer poly(ethylene glycol)-*block*-poly(propylene glycol)-*block*-poly(ethylene glycol) (EO₂₀PO₇₀EO₂₀) was dispersed in 30 g of water and 120 g of 2 M HCl solution and stirred for 5 h at room temperature. Subsequently, 9.5 g of tetraethyl orthosilicate (TEOS) was added drop-wise to the homogeneous solution at room temperature under constant stirring. The resulting gel was aged at 40 °C for 24 h and finally heated at 150 °C for 48 h. The solid was separated by filtration and dried at 80 °C overnight. Calcination was performed by heating the obtained powder material in air at 200 °C and 400 °C for 3 h at both temperatures and with the heating rate of 1 °C/min. Finally the temperature was increased to 550 °C and kept for 12 h in air to fully decompose the tri-block copolymer.

2.7. Characterization of mesoporous materials

Powder X-ray diffraction (XRD) patterns of the ordered mesoporous materials were collected on a Siemens D 5000 instrument using CuK α radiation (30 mA, 40 kV). The step width and time amounted to 0.02° and 3 s, respectively.

The textural properties of the samples, i.e., the specific BET surface areas, BJH pore volumes and pore size distributions were calculated from the nitrogen adsorption isotherms recorded at –196 °C using a Micromeritics ASAP 2010 equipment. Before the sorption measurements, the solid samples were pre-treated at 300 °C for unloaded samples and at 40 °C for samples loaded with enzyme under vacuum (<10^{–2} mbar) for 12 h.

2.8. Modelling P450 BM-3 heme domain and MD simulation

To obtain a homology model of the P450 BM-3 heme domain, two crystal structures were used as templates: the P450 BM-3 heme domain and a fusion protein between heme domain and FMN-domain [21,22]. The template structures were obtained from the Protein Data Bank (PDB entries 1BU7 and 1BVY). The N-terminal His6 tag, the C-terminal loop from residue 456 to 472, and the mutation F87A were modelled using MODELLER [23].

The initial model structure was refined by energy minimization, and a molecular dynamics (MD) simulation [24] of 600 ps was performed using the SANDER module of the AMBER 9 program suite [25]. The all-atom force field ff99 [26] was used, including the heme force field modification [27]. Prior to the simulation, the system was minimized using the steepest descent algorithm for 500 steps and then switching to the conjugant gradient algorithm for 1500 steps. The MD simulation was performed in implicit solvent by applying the generalized Born solvation model [28]. In the equilibration phase of 20 ps, a restraint force of 0.1 kcal/Å² mol was applied to the main chain atoms and a time step of 1 fs, a constant temperature of 300 K, and a cut-off value of 12 Å were used. After the equilibration, a MD simulation of 600 ps was performed with a time step of 2 fs and a maximum cut-off of 999 Å to ensure that it is larger than the protein size. To preserve the backbone structure a restraint force of 0.1 kcal/Å² mol was applied to main chain atoms of residues 36–493. For constraining the bond length involving hydrogen atoms the SHAKE algorithm [29] was applied with the default tolerance of 0.00001 Å and bond interactions involving hydrogen atoms were omitted.

For the visualization of the structures and the trajectories, Visual Molecular Dynamics (VMD) [30] and PyMOL [31] were used. To determine the molecular size of the protein, the minimal and maximal coordinates of the structure were calculated and a bounding box was constructed. The width of the bounding box corresponds to the maximum diameter of the protein.

2.9. Calculation of the titration curve and the electrostatic potential

For the calculation of the titration curve and the electrostatic potential, the complete sequence of BM3H.F87A was considered. The titration curve of the protein including the heme was calculated by MCCE [25,32] with a dielectric constant of the protein and the solvent of 8 and 80, respectively. Electrostatics was calculated by DELPHI using the finite Poisson–Boltzmann procedure [30] as implemented in the MCCE method. PARSE atomic charges and radii were used [27]. To visualize the charge distribution of P450 BM-3, the electrostatic potential was mapped on the structure using PyMOL.

3. Results and discussion

3.1. Characterization of the support materials

P450 BM-3 hydroxylates long-chain fatty acids in the presence of molecular oxygen and the cofactor NADPH. Various mutants of P450 BM-3 were shown to accept a broad range of substrates [33–37]. The heme domain of P450 BM-3 was found to catalyze the same reactions as the holoenzyme in the presence of hydrogen peroxide instead of dioxygen and the costly cofactor NADPH by following the so-called “peroxide shunt” [34,36]. As reported by Li and co-workers the replacement of phenylalanine at position 87 by alanine offered a higher peroxygenase activity of the BM-3 heme domain compared to the wild type enzyme [36]. The heme domain of P450 BM-3 F87A (further referred to as BM3H.F87A) was constructed as described in Section 2 and immobilized on three different silicates. The physical properties of the solid supports have a large influence on immobilization. Specific surface area, average pore diameter, and total specific pore volume were calculated from nitrogen adsorption isotherms and are collected in Table 1. Silica gel (Type 62) has unordered pores of non-uniform diameter ranging from 20 to 250 Å (mean pore diameter: 114 Å) and a specific surface area of ca. 327 m²/g. This material was used in order to realize the effect of the structure and pore diameter of enzyme loading capacity.

MCM-41 and SBA-15 are two commonly used ordered mesoporous molecular sieves which are characterized by channels of uniform dimension arranged in a strictly regular, parallel, non-intersecting hexagonal manner [38,39]. Generally, SBA-15 possesses pores with an adjustable uniform diameter between 60 Å and 150 Å, whereas MCM-41 typically has a pore diameter of approximately 20–60 Å. Besides that, SBA-15 has thick hydrothermal stable silica walls, different from MCM-41 which has thinner walls. The MCM-41 material used in this study has a pore diameter of 25 Å and a specific surface area of 1290 m²/g; SBA-15 was prepared with two differently large pore diameters of 60 Å and 133 Å with the corresponding specific surface areas of 828 and 380 m²/g. All three materials exhibit the well known type IV adsorption isotherms (Brunauer definition). As expected, no sorption hysteresis was observed for MCM-41, while a pronounced sorption hysteresis was found for the two SBA-15 materials. The sorption isotherms for the two SBA-15 materials before and after loading with enzyme are

Table 1
Physicochemical properties of the mesoporous materials used in this study.

Material	Pore diameter (Å)	BET surface area (m ² /g)	Total pore volume (cm ³ g ⁻¹)
MCM-41	25	1290	0.99
SBA-15	60	828	0.92
SBA-15	133	380	1.26
Silica gel	20–250	327	1.27

shown in Figs. 3 and 5 and will be discussed later. Moreover, the XRD-patterns show the characteristic reflections for ordered mesoporous materials with hexagonally arranged pores. Results of XRD characterization of the samples will also be discussed below.

3.2. Optimization of the immobilization procedure

The first set of experiments was conducted under various conditions in order to optimize the immobilization procedure. 20 mg of a mesoporous material were added to a P450 solution with initial concentration of 32 μM, mixed together and incubated as outlined below. The P450 loading at any particular time was calculated by centrifuging the solid material and measuring the P450 concentration of the supernatant, and taking these values from the initial P450 concentration, respectively.

The stability experiments with BM3H.F87A under the experimental conditions chosen for the immobilization processes revealed no loss in enzyme activity or solubility under stirring or slow rotation at different pH values and 10 °C or room temperature within 24 h. However, activity was completely lost upon intensive mixing in a bead mill already after 1 h. Therefore, mixing in a bead mill was excluded from the following experiments.

We observed, that agitation conditions influenced enzyme loading on both MCM-41 and SBA-15 (60 Å). Independently on immobilization time, pH and initial P450 concentration, a loading capacity of only 7–14 mg/g could be achieved after stirring. Under slow rotation the enzyme loading increased up to 30–34 mg/g for SBA-15 with a pore diameter of 60 Å and up to 22–26 mg/g for MCM-41 (25 Å), when immobilization was performed from a solution volume of 1.5 mL. For further experiments slow rotation was chosen as the most appropriated method. Our experiments demonstrated that under rotation in 1.5 mL volume equilibrium between enzyme solution and the solid materials as assessed from a constant loading was achieved already after 2 h.

The effect of pH on enzyme loading capacity was investigated using buffered enzyme solutions with different pH values ranging from pH 6.0 to pH 8.0 (supplementary Fig. 1). The highest enzyme loading after a single immersion of 33 mg P450 per 1 g SBA-15 (60 Å), 27 mg P450 per 1 g MCM-41 and 10 mg P450 per 1 g silica gel was observed at pH 7.0 (Fig. 1).

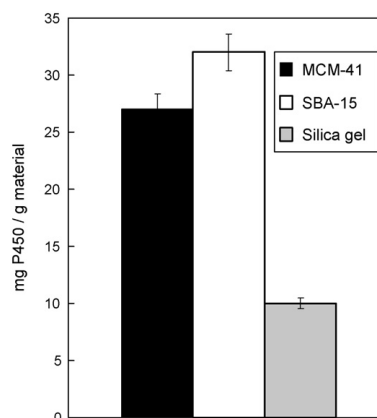
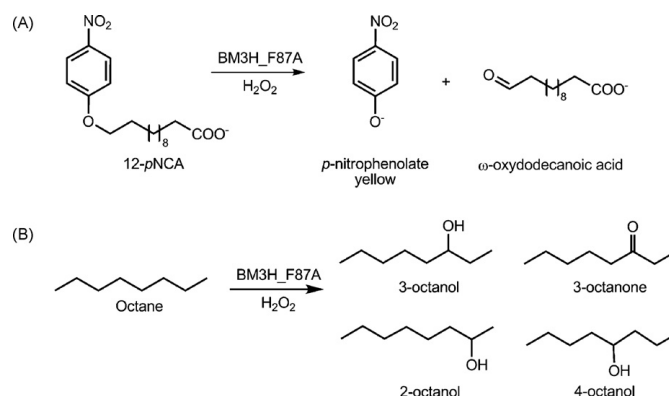


Fig. 1. Immobilization of BM3H.F87A on MCM-41 (25 Å), SBA-15 (60 Å) and silica gel after a single immersion at pH 7.0; enzyme concentration 35 μM, volume 1.5 mL.



Scheme 1. Hydroxylation of (A) 12-pNCA to *para*-nitrophenolate and (B) of *n*-octane to a mixture of secondary alcohols catalyzed by BM3H.F87A.

3.3. Influence of pore diameter on enzyme loading and activity

The dimensions of BM3H.F87A were estimated based on the crystal structure (PDB entries 1BU7 and 1BVY) after addition of the linker region and His6- tag in an extended conformation. During a 600 ps MD simulation their conformation changed to a more compact structure. The size of the modelled protein is $80 \times 70 \times 60 \text{ \AA}$, slightly larger than the crystal structure without the linker and His6- tag ($70 \times 60 \times 60 \text{ \AA}$). The distance along the minor axis (60 \AA) is similar to the pore diameter of SBA-15 (60 \AA), while the major axis exceeds the pore diameter by 20 nm. Thus, it is expected that the protein is able to bind inside the pores of silica gel and SBA-15, but not MCM-41 [12].

The preliminary experiments described above demonstrated the effect of pore diameter on enzyme loading capacity. We suggest that the low immobilization capacity of silica gel of $<10 \text{ mg/g}$ (measured after washing) relates to a broad distribution of pore diameters with an average at 114 \AA , which is significantly larger than the diameter of the enzyme. In this case up to 60% of P450 activity was removed from the support during washing as the enzyme molecules are probably only loosely adsorbed onto the internal surface of the silica gel. Almost no changes in the N_2 -adsorption isotherms or pore size distribution curves before and after loading BM3H.F87A on silica gel were observed (see supplementary Fig. 2).

The enzyme loading on ordered MCM-41 with a pore diameter of 25 \AA after a single immersion measured after repeating washing (15–20% P450 lost) was higher than that of sol-gel (27 mg/g vs. 10 mg/g). For an activity test, oxidation of *p*-nitrophenoxydodecanoic acid (12-pNCA) was chosen as a model reaction, because it allows the simple photometrical measurement of *p*-nitrophenolate which is produced during the reaction (Scheme 1A). The specific activity of the enzyme of $11 \text{ nmol product/mg P450/min}$ was very low compared to $1100 \text{ nmol product/mg P450/min}$ measured with the free enzyme. According to the literature data most proteins of molecular mass higher than 40 kDa cannot enter the pores of MCM-41 with pore diameters in the range of $20\text{--}30 \text{ \AA}$ and are only adsorbed on the external surface [6]. If enzyme is only immobilized on the external surface, the observed specific activity is usually low and enzyme leaching occurs. Our results with BM3H.F87A as well as physico-chemical analyses by N_2 -adsorption confirmed this observation. After loading of MCM-41 with the enzyme, no pore volume could be detected any more. However, the absence of any reflections in the X-ray diffractogram (Fig. 2A) indicates, that the long-range order of the MCM-41 was lost. To test whether the loss of the nitrogen sorption capacity is due to a pore blocking, the enzyme-loaded MCM-41 was re-calcined at $540 \text{ }^\circ\text{C}$ in air, but neither the XRD pattern typical for MCM-41-materials nor the nitrogen sorption capacity were restored. We therefore conclude that the loss of nitrogen sorption

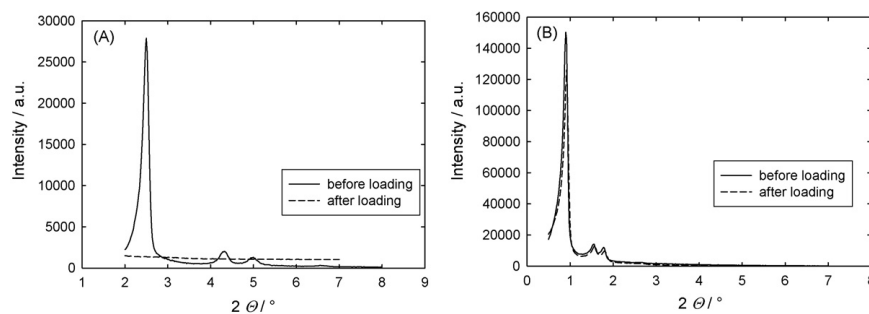


Fig. 2. XRD powder patterns of the mesoporous sieves before and after enzyme immobilization at pH 7.0: (A) MCM-41 (25 \AA); (B) SBA-15 (60 \AA).

Table 2

Loading capacity of the mesoporous materials and specific activity of immobilized BM3H.F87A toward 12-pNCA.

BM3H.F87A	Loading capacity (mg P450/g material)	Specific activity (nmol product/mg P450/min)	Immobilization procedure
Free enzyme	–	1100 ± 33	–
MCM-41 (25 Å)	27 ± 3	11 ± 2	Single immersion
SBA-15 (60 Å)	32 ± 4	112 ± 8	Single immersion
SBA-15 (60 Å)	50 ± 3	130 ± 10	Sequential immersing
SBA-15 (133 Å)	90 ± 6	830 ± 28	Single immersion

capacity could be due to a partial or total collapse of the ordered mesopore system in MCM-41 while loading with the enzyme. Note that the long-range order of SBA-15 (60 Å) remains intact even after loading of this support with the enzyme (Fig. 2B). This finding can be explained by the higher wall thickness of SBA-15 compared to MCM-41, and the resulting higher stability of SBA-15 toward the enzyme-containing solution during the immobilization steps.

Several reports suggest that for higher activity and stability the protein of interest should be immobilized inside the pores. However, the question whether the pore diameter should be significantly larger than the protein to allow for diffusion of protein and substrate into the pore or if pore and protein should be of similar size to increase stability and protection of the protein is under discussion, which is reflected by very contradictory reports published in this respect [40]. Takahashi et al. investigated immobilization of the horseradish peroxidase and reported that for its enhanced activity and stability in organic solvents, pores for the immobilization should match the size of the protein, because, if the pore will be too big, the enzyme will not be well protected [41,42]. Our investigations demonstrated that although the immobilization capacity of SBA-15 (60 Å) was similar to MCM-41 (25 Å) (25–32 mg/g) after a single immersion, the observed specific activity of the immobilized enzyme toward 12-pNCA was at least 10-fold higher (11 nmol product/mg P450/min vs. 112 nmol product/mg P450/min) (Table 2). This is a first hint for the accommodation of BM3H.F87A molecules inside the pores of SBA-15. Our results demonstrate that by matching pore size with protein size higher enzyme activity can be attained. However, since almost the same amount of P450 was immobilized on SBA-15 (60 Å) and MCM-41 (25 Å) under the same conditions, the loading is obviously further influenced by the available pore volume and surface area.

Another indication of adsorption of the enzyme within the pores of SBA-15 comes from the comparison of the nitrogen sorption isotherms of this material before and after loading with enzyme and that of the amorphous silica gel. In SBA-15 (60 Å), the specific surface area was reduced from 828 m²/g before to 539 m²/g after loading with enzyme (Fig. 3). At variance, the specific surface area was reduced from 327 to only 302 m²/g for silica gel. Evi-

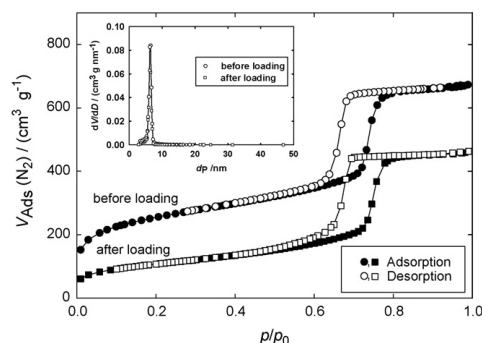


Fig. 3. Changes of N₂-adsorption isotherms ($T = -196^\circ\text{C}$) and pore size distribution curves from the adsorption (●, ■) and desorption (○, □) before (○) and after (□) loading BM3H.F87A on SBA-15 (60 Å).

dently, a lower amount of enzyme was adsorbed on silica gel, where only surface area within rather large pores is available and enzyme removal during washing is facilitated. An at least partial immobilization of the enzyme within the pores of SBA-15 is therefore likely. Also note that the lower closure point of the hysteresis loop for the SBA-15 after enzyme immobilization (Fig. 3) is shifted to a relative pressure <0.6 which might be interpreted in terms of unevenly shaped pores due to presence of enzyme on the inner surface.

3.4. Optimization of enzyme loading onto SBA-15

In an attempt to improve immobilization onto SBA-15 (60 Å), we increased the initial P450 concentration in solution. Generally, after a simple immersion in a total volume of 1.5 mL, the loading capacity of SBA-15 increased only until a BM3H.F87A-concentration of 20 μM (Fig. 4A). Further increase up to 50 μM P450 did not change loading capacity of SBA-15 (60 Å). Furthermore, P450 concentrations higher than 50 μM limited somehow the immobilization process, resulting even in reduced loading capacity. 25 mg were loaded on 1 g SBA-15 when 31 μM BM3H.F87A was used, only 17 mg/g with 75 μM, and enzyme concentration of 112 μM led to the lowest enzyme loading of 10 mg/g. The calculation of the inner pore volume (Table 1) suggested that in all cases it was high enough for immobilization of at least a 50-fold higher P450 concentration. According to the modelling studies it is expected that the protein is able to bind inside the pores of SBA-15 (60 Å). However, the observed behaviour indicates the existence of some sort of diffusion hindrance or pore blockage which prevents the protein molecules from diffusing into the inner particle region. Remarkably, this hindrance could be overcome by sequen-

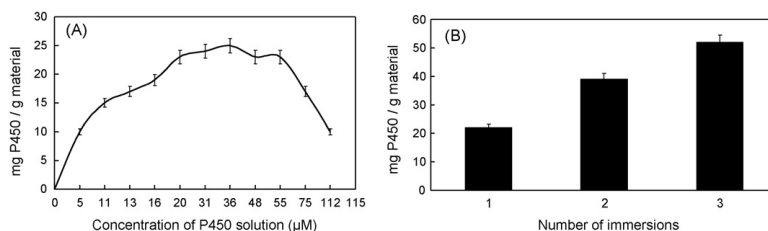


Fig. 4. Optimization of enzyme loading onto SBA-15; (A) Dependence of the immobilization efficiency of SBA-15 (60 Å) on initial enzyme concentration; (B) Sequential immersion of SBA-15 (60 Å) at enzyme concentration of 36.4 μM.

tial immersing the same portion of supporting material with fresh enzyme solution. 20 mg SBA-15 was loaded with a 36.4 μM P450 solution several times in 1.5 mL under equal conditions (Fig. 4B). After the first round of immobilization 22 mg/g was loaded onto SBA-15. More than 90% of P450 was immobilized from the loading solution. After several washing steps the same material was loaded again, which led to an increased loading capacity up to 37 mg/g. Only 68% of the enzyme was adsorbed during this second immersion round. The next immersion resulted in totally 50 mg/g, however only 35% of P450 was immobilized in this case. Nevertheless, also in this case the specific activity of BM3H.F87A retained on the similar level and achieved 130 nmol/mg P450/min (Table 2).

The fact that reloading with fresh enzyme solutions increased the amount of the absorbed protein, and that high concentrations of the initial P450 solution resulted in a lower immobilization efficiency indicated, that loading might be kinetically controlled. Our further experiments demonstrated that increase of immobilization volume from 1.5 up to 10 mL represents an alternative way for improving the loading capacity of SBA-15 at P450 concentrations of $>50 \mu\text{M}$ or higher. In 3 mL of 48 μM BM3H.F87A an almost three-fold increase in loading capacity (75 mg/g) was achieved. When the same experiment was performed in 10 mL, loading capacity reached 180 mg/g within 2 h (data not shown). These results confirmed the presence of diffusion hindrance or pore blockage which prevents the protein molecules from diffusing into the inner particle region.

The N_2 -adsorption isotherms estimated before and after immobilization of BM3H.F87A onto SBA-15 (60 Å) demonstrated a reduction of pore volume from 828 to 539 m^2/g (see above). However, as discussed earlier, such a finding does not necessarily give absolute evidence for entrance of the enzyme into the pores, as pore blockages at the entrance to the mesopores can reduce pore volumes and surface areas even when the protein molecules have not fully entered the pore [43–45]. Furthermore, since the specific activity of BM3H.F87A immobilized on SBA-15 with a pore diameter of 60 Å was still lower compared to its free, i.e., unsupported form, we tested SBA-15 with a pore diameter of 133 Å. This diameter is twice as large as the longest axis of BM3H.F87A and should be enough for “in-pore” protein accommodation. As expected, the loading capacity of SBA-15 (133 Å) was higher and reached 50 mg/g when the experiment was carried out in 1.5 mL of 26.8 μM P450 solution. Moreover, no P450 was removed during the washing steps. In 10 mL of 16.2 μM P450 solution 90 mg active enzyme could be immobilized onto 1 g SBA-15 and increased up to 210 mg/g SBA-15 after two additional immersions. For identification of enzyme leaching under reaction conditions the reaction mixture (without the loaded materials) was analysed after 3 and 5 min of p-NCA oxidation as described in Section 2. Neither spectral data nor activity measurements indicated the presence of the P450 monooxygenase in the reaction mixture within the first 5 min of reaction.

Upon loading with enzyme, the specific surface area of the SBA-15 (133 Å) decreased from 380 to 268 m^2/g . Although this decrease is less pronounced than that for the SBA-15 with the smaller pore diameter of 60 Å, it is still remarkable. Moreover, in the case of SBA-15 (133 Å) a significant change of the pore diameter distribution was observed (Fig. 5). While the pore size distribution was generally broadened, the most apparent effect is that a larger amount of smaller pore diameters than in the enzyme-free support material is observed. Also, the maximum of the pore size distribution was shifted from 133 to 117 Å. Taken together, the reduction in specific surface area and the shift of the pore size distribution to significantly lower values can be considered as evidence for an immobilization of BM3H.F87A within the mesopores of SBA-15 (133 Å).

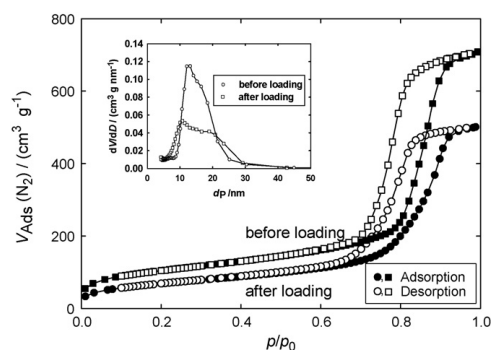


Fig. 5. Changes of N_2 -adsorption isotherms ($T = -196^\circ\text{C}$) and pore size distribution curves from the adsorption (●, ■) and desorption (○, □) before (○) and after (□) loading BM3H.F87A on SBA-15 (133 Å).

Additionally, the specific activity of the immobilized BM3H.F87A as measured with 12-pNCA was tested. For the enzyme supported on SBA-15 (133 Å) it increased up to 840 nmol product/mg P450/min (Table 2). As mentioned in the introduction, previously we immobilized a triple P450 BM-3 mutant in a sol-gel matrix derived from tetraethoxyorthosilicate. In that study the natural holoenzyme, consisting of a monooxygenase domain fused to a reductase domain was used. As electron donor the cofactor NADPH, supported by regeneration with formate dehydrogenase was applied. Under these conditions the specific activity of the immobilized enzyme toward 12-pNCA reached 884 nmol/mg P450/min (in the original manuscript expressed as 0.89 U/mg P450). This value is comparable with activity obtained with the enzyme immobilized on SBA-15 (133 Å). One should take into account that the molecular mass of the holoenzyme is two-fold higher than that of the separated monooxygenase.

Nevertheless, the obtained value was still lower compared to the free enzyme (1100 nmol/mg P450/min). As possible reasons for this difference in specific activity (1) a non-optimal orientation of enzyme molecule in the mesopores of the support, and (2) insufficient accessibility of the substrate 12-pNCA for immobilized BM3H.F87A, can be considered.

3.5. Modelling the P450 BM-3 heme domain

To explain the observed results and to elucidate enzyme orientation on SBA-15, the electrostatic properties of the protein and the immobilization matrix were considered. It was assumed that long-range electrostatic interactions are dominating the pH-dependent interactions. Because the pore size exceeds the diameter of the protein, the matrix was modelled as a planar surface. Thus, a titration curve of BM3H.F87A was calculated for pH values between 3 and 10. The total charge of the protein was decreasing from 32 to -18 with increasing pH, with a pI of 5.4. The electrostatic potential was calculated for pH 6.0, 7.0, and 8.0 (Fig. 6). At pH 7.0 where the maximal amount of BM3H.F87A was immobilized, the protein is negatively charged (total charge of -7.9). The electrostatic surface of BM3H.F87A consists of large patches of negative charge with a small patch of positive charge proximal to the heme group (Fig. 6). At this pH, the surface of the matrix is negatively charged, since the point-of-zero of aluminium silica is approximately at pH of 3.0 [46]. Previously, it has been shown that charged proteins bind preferably to a surface with the opposite charge, less well to a surface with the same charge, and least to an uncharged surface [47]. It has been suggested that binding occurs via charged surface patches,

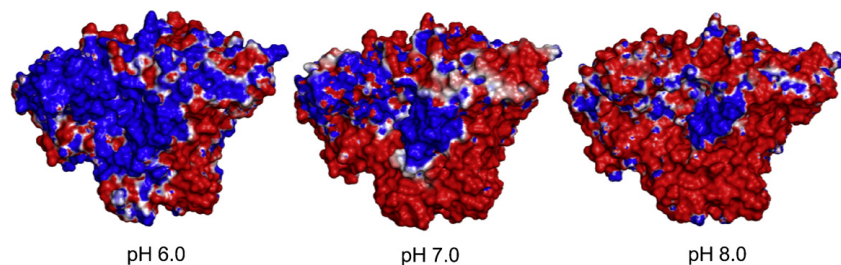


Fig. 6. Electrostatic potential at pH 6.0, 7.0 and 8.0 mapped on the surface of the structure model of P450 BM-3 heme domain. A positive surface potential is shown in blue and a negative surface potential is shown in red.

thus the orientation of a protein binding to a positively or a negatively charged surface is expected to change [47]. This observation could explain the peak in the immobilization efficiency at pH 7.0. At this pH, the matrix is slightly charged, and the positive patch of P450 is large enough to enable binding, despite of the repulsion between the negatively charged surface and the negative charge of the protein. At a higher pH value, however, the negative charge of the protein and the surface increase, and the size of the positive patch of the protein decreases, which leads to a decrease of immobilization efficiency. For pH values below 7.0, the charge at the surface decreases, which leads to a decrease of immobilization efficiency, as the surface becomes neutral. In addition, at pH 7.0 the protein binds in the favourable orientation, since the positive patch is located proximal to the heme at the reductase binding site, opposite to the entrance to the substrate binding site. This suggests that substrate binding pocket of the enzyme should be accessible for substrates at least in SBA-15 with pore diameter of 133 Å, and so cannot be a limiting factor for specific activity.

3.6. Activity and stability of immobilized BM3H.F87A toward *n*-octane

In our previous work we observed that the substrate 12-*p*NCA as well as its oxidation products can bind to negatively charged supporting materials like DEAD-cellulose. In order to elucidate whether either *p*NCA or the *p*-nitrophenolate bind to SBA-15 and this can affect the activity of the immobilized P450, hydrophobic *n*-octane was used as substrate. The measurements with purified free BM3H.F87A as well as with immobilized enzyme were carried out for 2 h under continuous shaking at room temperature.

Conversion of *n*-octane reached 18–20% with immobilized BM3H.F87A and was even higher than in the system with free enzyme (9%). As reaction products the regioisomers 2-, 3-, 4-octanol and 3-octanone were identified in molar ratio of approx 2: 4: 3: 1 (Scheme 1B). These values correlate to those reported previously [33] and indicate that the regioselectivity of BM3H.F87A was not changed upon immobilization.

Generally the observed total P450 activity for both immobilized and free enzyme was quite low and reached 35–60 nmol total product/mg P450 (Table 3). As *n*-octane oxidation by BM3H.F87A

Table 3
Conversion of *n*-octane by free and immobilized BM3H.F87A after 2 h. Identical amounts of free and immobilized enzyme (3 mg) and 10 μM H₂O₂ were used in all experiments.

Enzyme	Substrate conversion (%)	Activity (nmol total product/mg P450)
Free enzyme	9	30 ± 4
SBA-15 (60 Å)	20	62 ± 6
SBA-15 (133 Å)	18	59 ± 7

is much slower than the oxidation of 12-*p*NCA, the enzyme stability in the presence of 10 mM H₂O₂ becomes a limiting factor for conversion. Judging from the obtained conversion values we suggest, that the enzyme hidden inside the pores stayed stable over at least two-times longer period of time and produced more oxidized products than free enzyme under the same conditions. Since conversion values for SBA-15 with different pore diameters were similar, obviously *n*-octane can reach the immobilized P450 even inside the smaller pores. This may indicate that lower activity of BM3H.F87A immobilized onto SBA-15 (133 Å) toward 12-*p*NCA compared to its free form might be due to insufficient accessibility of the substrate 12-*p*NCA for the enzyme. However, deactivation of the enzyme inside the pores cannot be excluded completely, since this parameter cannot be elucidated.

4. Conclusions

Efficient immobilization of P450 monooxygenases can only be achieved when a detailed understanding of the enzyme properties is combined with a tailoring design of mesoporous supports. Our results on the immobilization of the heme domain BM3H.F87A on mesoporous molecular sieves MCM-41 (25 Å) and SBA-15 (60 Å and 133 Å) demonstrated the importance of a match between pore diameter and protein dimensions. The enzymatic activity was retained only after immobilization of BM3H.F87A inside the pores of SBA-15. If the enzyme was adsorbed on the external surface of the material as in the case of MCM-41, its activity was very low. Furthermore, for the “in pore” immobilized BM3H.F87A the nature of the substrate plays a critical role. Reduced accessibility of 12-*p*NCA for the immobilized BM3H.F87A is probably a limiting factor, which leads to reduction in enzyme activity. For *n*-octane, the operational stability of the enzyme becomes a more essential issue, reflecting the fact that the immobilized enzyme showed higher activity than its free form.

Acknowledgements

This work was financially supported by the Deutsche Forschungsgemeinschaft (SFB 706), the Ministerium für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg and the Fonds der Chemischen Industrie. We acknowledge valuable contributions by Alexander Steudle.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molcatb.2010.01.020.

References

- [1] D.L. King, M.R. Azari, A. Wiseman, *Methods Enzymol.* 137 (1988) 675–686.
- [2] S.B. Lamb, D.C. Lamb, S.L. Kelly, D.C. Stuckey, *FEBS Lett.* 431 (1998) 343–346.
- [3] M. Taylor, D.C. Lamb, R.J. Cannell, M.J. Dawson, S.L. Kelly, *Biochem. Biophys. Res. Commun.* 279 (2000) 708–711.
- [4] S. Maurer, V. Urlacher, H. Schulze, R.D. Schmid, *Adv. Synth. Catal.* 345 (2003) 802–810.
- [5] H. Li, T.L. Poulos, *Nat. Struct. Biol.* 4 (1997) 140–146.
- [6] J.F. Diaz, K.J. Balkus, *J. Mol. Catal. B: Enzym.* 2 (1996) 115–126.
- [7] C.H. Lee, J. Lang, C.W. Yen, P.C. Shih, T.S. Lin, C.Y. Mou, *J. Phys. Chem. B* 109 (2005) 12277–12286.
- [8] A. Katiyar, S. Yadav, P.G. Smirniotis, N.G. Pinto, *J. Chromatogr. A* 1122 (2006) 13–20.
- [9] E.L. Pires, E.A. Miranda, G.P. Valenca, *Appl. Biochem. Biotechnol.* 98 (2002) 963–976.
- [10] A. Salis, D. Meloni, S. Ligas, M.F. Casula, M. Monduzzi, V. Solinas, E. Dumitriu, *Langmuir* 21 (2005) 5511–5516.
- [11] S.W. Song, K. Hidajat, S. Kawi, *Langmuir* 21 (2005) 9568–9575.
- [12] M.C.R. Hernandez, J.E.M. Wejbe, J.I.V. Alcantara, R.M. Ruvalcaba, L.A.G. Serrano, J.T. Ferrara, *Microporous Mesoporous Mater.* 80 (2005) 25–31.
- [13] M. Rosales-Hernandez, L. Kispert, E. Torres-Ramirez, D. Ramirez-Rosales, R. Zamorano-Ulloa, J. Trujillo-Ferrara, *Biotechnol. Lett.* 29 (2007) 919–924.
- [14] U. Schwaneberg, C. Schmidt-Dannert, J. Schmitt, R.D. Schmid, *Anal. Biochem.* 269 (1999) 359–366.
- [15] S. Pflug, S.M. Richter, V.B. Urlacher, *J. Biotechnol.* 129 (2007) 481–488.
- [16] T. Omura, R.J. Sato, *J. Biol. Chem.* 239 (1964) 2370–2378.
- [17] S.C. Maurer, K. Kuehnel, L.A. Kaysser, S. Eiben, R.D. Schmid, V.B. Urlacher, *Adv. Synth. Catal.* 347 (2005) 1090–1098.
- [18] T. Boger, R. Roesky, R. Glaser, S. Ernst, G. Eigenberger, J. Weitkamp, *Microporous Mater.* 8 (1997) 79–91.
- [19] M. Choi, W. Heo, F. Kleitz, R. Ryoo, *Chem. Commun. (Camb)* (2003) 1340–1341.
- [20] A. Vinu, V. Murugesan, O. Tangermann, M. Hartmann, *Chem. Mater.* 16 (2004) 3056–3065.
- [21] I.F. Sevioukova, H.Y. Li, H. Zhang, J.A. Peterson, T.L. Poulos, *Proc. Natl. Acad. Sci. U.S.A.* 96 (1999) 1863–1868.
- [22] H.M. Berman, T. Battistuzzi, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, C. Zardecki, *Acta Crystallogr. D. Biol. Crystallogr.* 58 (2002) 899–907.
- [23] A. Sali, T.L. Blundell, *J. Mol. Biol.* 234 (1993) 779–815.
- [24] K.K. Khan, S. Mazumdar, S. Modi, M. Sutcliffe, G.C. Roberts, S. Mitra, *Eur. J. Biochem.* 244 (1997) 361–370.
- [25] R.E. Georgescu, E.G. Alexov, M.R. Gunner, *Biophys. J.* 83 (2002) 1731–1748.
- [26] J.M. Wang, P. Cieplak, P.A. Kollman, *J. Comput. Chem.* 21 (2000) 1049–1074.
- [27] D. Sitkoff, K.A. Sharp, B. Honig, *J. Phys. Chem. B* 98 (1994) 1978–1988.
- [28] C.J. Cramer, D.G. Truhlar, *Chem. Rev.* 99 (1999) 2161–2200.
- [29] R.C. Walker, M.F. Crowley, D.A. Case, *J. Comput. Chem.* 29 (2008) 1019–1031.
- [30] P.A. Zoretic, H. Fang, A.A. Ribeiro, *J. Org. Chem.* 63 (1998) 7213–7217.
- [31] W.L. Delano, in: San Carlos, CA, USA: Delano Scientific, 2002.
- [32] E.G. Alexov, M.R. Gunner, *Biophys. J.* 72 (1997) 2075–2093.
- [33] D. Appel, R.D. Schmid, C.A. Dragan, M. Bureik, V.B. Urlacher, *Anal. Bioanal. Chem.* 383 (2005) 182–186.
- [34] P.C. Cirino, F.H. Arnold, *Angew. Chem. Int. Ed. Engl.* 42 (2003) 3299–3301.
- [35] M. Landwehr, L. Hochrein, C.R. Otey, A. Kasrayan, J.E. Backvall, F.H. Arnold, *J. Am. Chem. Soc.* 128 (2006) 6058–6059.
- [36] Q.S. Li, J. Ogawa, S. Shimizu, *Biochem. Biophys. Res. Commun.* 280 (2001) 1258–1261.
- [37] V.B. Urlacher, A. Makhsumkhanov, R.D. Schmid, *Appl. Microbiol. Biotechnol.* 70 (2006) 53–59.
- [38] C.T. Kresge, M.E. Leonowicz, W.J. Roth, J.C. Vartuli, J.S. Beck, *Nature* 359 (1992) 710–712.
- [39] M. Kruk, M. Jaroniec, C.H. Ko, R. Ryoo, *Chem. Mater.* 12 (2000) 1961–1968.
- [40] S. Hudson, J. Cooney, E. Magner, *Angew. Chem. Int. Ed. Engl.* 47 (2008) 8582–8594.
- [41] H. Takahashi, B. Li, T. Sasaki, C. Miyazaki, T. Kajino, S. Inagaki, *Chem. Mater.* 12 (2000) 3301–3305.
- [42] H. Takahashi, B. Li, T. Sasaki, C. Miyazaki, T. Kajino, S. Inagaki, *Microporous Mesoporous Mater.* 44 (2001) 755–762.
- [43] J. Deere, E. Magner, J.G. Wall, B.K. Hodnett, *J. Phys. Chem. B* 106 (2002) 7340–7347.
- [44] A. Vinu, V. Murugesan, M. Hartmann, *J. Phys. Chem. B* 108 (2004) 7323–7330.
- [45] A. Vinu, C. Streb, V. Murugesan, M. Hartmann, *J. Phys. Chem. B* 107 (2003) 8297–8299.
- [46] A.J. O'Connor, A. Hokura, J.M. Kisler, S. Shogo, G.W. Stevens, Y. Komatsu, *Sep. Purif. Technol.* 48 (2006) 197–201.
- [47] H. Noh, S.T. Yohe, E.A. Vogler, *Biomaterials* 29 (2008) 2033–2048.

6.5 The Laccase Engineering Database: A Classification and Analysis System for Laccases and Related Multicopper Oxidases

Eingereicht bei *BMC Genomics*

Sirim, D., Wagner, F., Wang, L., Schmid, R. D., Pleiss, J., The Laccase Engineering Database: A Classification and Analysis System for Laccases and Related Multicopper Oxidases.

6.5.1 Abstract

Background

Laccases and their homologues form the protein superfamily of multicopper oxidases (MCO). They catalyze the oxidation of many, particularly phenolic substances, and, besides playing an important role in many cellular activities, are of interest in biotechnological applications.

Description

The Laccase Engineering Database (*LccED*, <http://www.LccED.uni-stuttgart.de>) was designed to serve as a tool for a systematic sequence-based classification and analysis of the diverse multicopper oxidase protein family. More than 2200 proteins were classified into 10 superfamilies and 55 homologous families. For each family, the *LccED* provides multiple sequence alignments, phylogenetic trees, and family-specific HMM profiles. The integration of structures for 14 different proteins allows a comprehensive comparison of sequences and structures to derive biochemical properties. Among the families, the distribution of the proteins regarding different kingdoms was investigated. The database was applied to perform a comprehensive analysis by MCO- and laccase-specific patterns.

Conclusions

The *LccED* combines information of sequences and structures of MCOs. It serves as a classification tool to assign new proteins to a homologous family and can be applied to investigate sequence-structure-function relationship and to guide protein engineering.

6.5.2 Background

Multicopper oxidases (MCOs) catalyze the one-electron oxidation of their substrates with a concomitant four-electron reduction of molecular oxygen to water. MCOs consist of four enzyme families: laccases (EC 1.10.3.2), ascorbate oxidases (EC 1.10.3.3), ferroxidases (EC 1.16.3.1), and ceruloplasmin (EC 1.16.3.1). Functional studies have revealed that MCOs have two active sites: one blue type 1 (T1) copper site where the substrate is oxidized, and a trinuclear copper cluster (consisting of three type 2 (T2)/ type 3 (T3) coppers) where oxygen is bound, activated, and reduced (Solomon et al., 1996). The electrons are transferred from the T1 site to the T2/T3 site via highly conserved amino acids which have previously been described in PROSITE (Hulo et al., 2008; Sigrist et al., 2002) as MCO-specific patterns, further referred to as M2 and M4 (Messerschmidt and Huber, 1990; Ouzounis and Sander, 1991). In addition, laccase-specific signature sequences, namely L1 and L3, were generated from 100 plant and fungal laccase sequences. L1 and L3 have been suggested to be specific for laccases and were proposed to distinguish laccases from other MCOs (Kumar et al., 2003). While there is only low overall sequence similarity, the structure and catalytic mechanism is conserved (Nakamura and Go, 2005). Most MCOs consist of three cupredoxin domains, except for ceruloplasmin and some bacterial laccases which contain six or two domains, respectively (Murphy et al., 1997). Depending on the number of domains, MCOs vary in size, from 300 to 1000 residues, and contain up to six copper ions (Messerschmidt and Huber, 1990).

Laccases, which constitute the largest subfamily of MCOs, are widely distributed among fungi, higher plants (Mayer and Staples, 2002; Messerschmidt, 1997), bacteria (Alexandre and Zhulin, 2000) and insects (Dittmer et al., 2004). In fungi, they are involved in lignin degradation (Bourbonnais and Paice, 1990), pigment production (Clutterbuck, 1972) and plant pathogenesis (Geiger et al., 1986). In plants, they mainly catalyze biosynthesis of lignin (O'Malley et al., 1993). In bacteria, they are suggested to play a role in melanin production, spore coat resistance, morphogenesis, and detoxification of copper (Sharma et al., 2007). In particular laccases which form the largest subgroup of MCOs, also have a high biotechnological potential as versatile catalysts in textile and in pulp and paper industries, as well as in food applications, bioremediation, and organic synthesis (Riva, 2006; Couto and Herrera, 2006). However, their selectivity and redox potential are often restricted. Engineered laccases promise to have improved enzymatic properties such as activity, specificity, and selectivity (Kunamneni et al., 2008). It is expected that, understanding the relationships between sequence, structure, and function would

greatly help the engineering of laccases. Therefore we integrated data on MCO sequences and structures and built up the Laccase Engineering Database (*LccED*) using the data warehouse system *DWARF* (Fischer et al., 2006). Previously, 350 MCOs were assigned to ten superfamilies (Hoegger et al., 2006): A) basidiomycete laccases, B) ascomycete laccases, C) insect laccases, D) fungal pigment MCOs, E) fungal ferroxidases, F) fungal and plant ascorbate oxidases, G) plant laccase-like MCOs, H) copper resistance proteins (CopA), I) bilirubin oxidases, and J) copper efflux (CueO) proteins. Homologous MCO sequences were retrieved and assigned to families by sequence similarity. In order to assist comprehensive sequence analysis, reliable multisequence alignments were generated and annotated either by an automated pattern search or by information extracted from GenBank (Benson et al., 2008). In addition, family-specific HMM profiles (Durbin et al., 1998) and a BLAST Altschul et al. (1997) interface are provided to allow an assignment of new sequences to families. Thus, the *LccED* is the first data resource that combines information on sequences, sequence alignments, annotations, and structures of MCOs.

6.5.3 Construction and content

Database construction

The *LccED* was established within the data warehouse system *DWARF*, which provides a data model for the integration of sequences and structures in a family-specific protein database, as well as tools for extracting and loading data from various data sources (Fischer et al., 2006). Previously, more than 350 MCO sequences were assigned to ten superfamilies (Hoegger et al., 2006). From this data set 248 sequences, for which a GenBank entry was available, were selected as seed sequences and assigned to the ten superfamilies, which were named based on the origin of their seed sequences. Subsequently, for each seed sequence a BLAST search (Altschul et al., 1997) was performed in the non-redundant sequence database at NCBI (<http://ncbi.nlm.nih.gov>) with an E-value of $E = 10^{-10}$. For the more diverse bacterial families a higher E-value of $E = 10^{-5}$ was applied. Each BLAST hit was assigned to the superfamily of the respective seed sequence if the sequence identity was higher than 40 %. Sequences within a superfamily were classified into homologous families based on multiple sequence alignments and phylogenetic trees, as calculated by CLUSTAL W (Thompson et al., 1994). Information on source organism, sequence annotations, and sequence was extracted by the sequence data loader of the *DWARF* system. The species information was adapted to the NCBI taxonomy. Different names denominating the same organism are listed as synonyms

on the organism page. Sequences from the same source organism and sharing > 98 % identical residues were represented as one single protein entry. This assignment is implemented in an automated script, thus preventing that one protein from the same organism may occur in duplicate within the database and avoiding redundancy even if it may occur in GenBank.

In case of BLAST hits specifying a protein structure, the respective structure was extracted from the PDB (Berman et al., 2002), stored as structural monomers, and secondary structure information was generated for each chain by DSSP (Kabsch and Sander, 1983). For all families, multiple sequence alignments were performed by CLUSTAL W (Thompson et al., 1994) and manually checked to improve consistency and quality. Proteins which were not assignable to any superfamily and obviously did not belong to the class of MCOs were removed from the database. The *LccED* is regularly updated using the automated update function of the *DWARF* system, keeping pace with the permanently growing GenBank data (Benson et al., 2008).

6.5.4 Contents

The *LccED* contains data on 2804 sequences and 2274 proteins. For 14 proteins from 6 different homologous families crystal structures are deposited, which results in a total of 68 structural monomers. The proteins were assigned to 10 superfamilies based on the origin of the seed sequences and to 55 homologous families based on phylogeny (table 6.5.1). For each superfamily and homologous family an annotated multiple sequence alignment, a phylogenetic tree, and a family-specific HMM profile (<http://hmmer.janelia.org/>) were generated.

Table 6.5.1: *LccED* families, sequences, and structures

<i>Superfamily</i>	<i>Homologous families</i>	<i>Proteins</i>	<i>Structures</i>
A (Basidiomycete Laccases)	4	201	13
B (Ascomycete Laccases)	6	421	6
C (Insect Laccases)	8	168	0
D (Fungal Pigment MCOs)	4	55	0
E (Fungal Ferroxidases)	6	117	6
F (Fungal and plant AOs)	6	137	8
G (Plant Laccases)	5	333	0
H (Bacterial CopA Proteins)	6	383	0
I (Bacterial Bilirubin Oxidases)	5	149	24
J (Bacterial CueO Proteins)	5	310	11

6.5.5 Utility and discussion

Web interface

The *LccED* is publicly available on <http://www.LccED.uni-stuttgart.de>. It can be browsed by family, organism, or structure. For each family, pre-calculated annotated multiple sequence alignments, phylogenetic trees, and HMMs are provided. All protein entries in the alignments and trees are linked to their original NCBI entries. Functionally relevant amino acids are colour coded, and further information is displayed upon moving the mouse over the respective residue in the multiple sequence alignment. The conservation degree of the alignment was calculated using PLOTCON (Rice et al., 2000). Phylogenetic trees are visualized by an in-house developed tree-visualizer which allows coloring each entry by properties such as homologous family (in superfamily-trees), organism, sequence length, and kingdom of the source organism (figure 6.5.2). Via a local BLAST interface, unknown MCO sequences can be classified by sequence similarity to the existing *LccED* entries. A tar archive comprising all information on families, sequences, structures, multiple sequence alignments, trees, and profiles can be downloaded.

Analysis of organism distribution and sequence patterns

In this study, 2274 MCO proteins from a wide spectrum of source organisms were assigned to superfamilies and homologous families based on sequence similarity and phylogenetic analysis. A comprehensive analysis of the relationships between sequence similarity, source organism and of patterns forming the binding sites of copper was performed in all 2274 proteins. The proposed patterns L1 ($H-W-H-G-x(9)-D-G-x(5)-Q-C-P-I$) and L3 ($H-P-x-H-L-H-G-H$) have been suggested to be specific for laccases, the patterns M2 ($G-x-[FYW]-x-[LIVMFYW]-x-[CST]-x-PR-K-x(2)-S-x-LFH-G-[LM]-x(3)-[LIVMFYW]$, PROSITE entry PS00079) and M4 ($H-C-H-x(3)-H-x(3)-[AG]-[LM]$, PROSITE entry PS00079) for MCOs (figure 6.5.1). Pattern L1 includes one histidine which binds the T2 copper and one histidine which binds the T3 copper. Pattern M2 includes two further T3 copper ligands. Pattern L3 includes ligands of the T1, T2, and T3 coppers. Within pattern M4 three of the four ligands of the T1 centre and one T3 copper ligand are located (figure 6.5.1). For annotation and evaluation purposes regular expressions generated from L1, L3, M2 and M4 were applied.

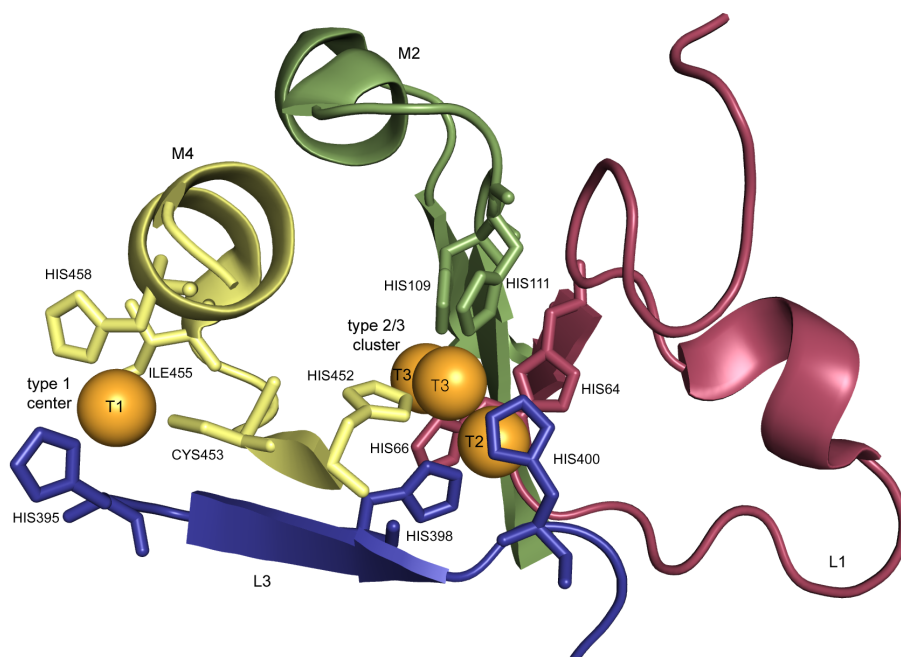


Figure 6.5.1: Copper binding residues of laccase from *T. versicolor* (PDB entry 1GYC (Piontek et al., 2002)). The copper centres are shown in orange, the residues that match the defined pattern L1, M2, L3, M4 are coloured in red, green, blue, and yellow respectively (visualization by PyMOL (Delano, 2002)).

Family **A** (Basidiomycete Laccases) contains exclusively fungal proteins, 91 % are from basidiomycetes (homologous families A1 - A4), 9 % from ascomycetes (homologous family A2). 80 % are annotated as laccases in GenBank. 23 % of the proteins contain the pattern L1, 22 % M2, 14 % L3, and 46 % M4.

Family **B** (Ascomycete Laccases) contains 36 % from ascomycetes. All of them cluster into the homologous family B1 and 62 % are annotated as laccases in GenBank. The other proteins are all of bacterial origin (homologous families B2 - B6) and 3 % are annotated as laccases in GenBank. Yet, they show a considerable sequence identity of over 40 % to ascomycetous laccases. 88% of the proteins contain the pattern L1, 92 % M2, 49 % L3, and 15 % M4.

Family **C** (Insect Laccases) resulted in 78 % proteins of insect origin (homologous families C1 - C8). The remaining 22 % consist of euechinoidea (in homologous family C1), cephalochordata and cnidaria (in homologous family C6). 38 % are annotated as laccases in GenBank. 30 % of the proteins contain pattern L1, 75 % M2, 75 % L3, and 3 % M4.

Family **D** (Fungal Pigment MCOs) contains exclusively fungal proteins. 36 % of the proteins are annotated in GenBank as laccases. 90 % of the proteins contain pattern L1, 78 % M2, 82 % L3, and 11 % M4.

Family **E** (Fungal Ferroxidases) contains exclusively fungal proteins. 17 % of the proteins are annotated in GenBank as laccases. 83 % of the proteins contain pattern L1, 40 % M2, 23 % L3, and 3 % M4.

Family **F** (Fungal and Plant Ascorbate Oxidases) mainly contain proteins of plant origin (homologous families F2 - F6). 12 % are of fungal origin and clustered all to the homologous family F1. 2 % are annotated as laccases in GenBank. In this family, 88 % of the proteins contain pattern L1, 56 % M2, 66 % L3, and 66 % M4.

Family **G** (Plant Laccases) exclusively contains proteins of plant origin and 83 % are annotated as laccases in GenBank (homologous families G1 - G5). 15 % of the proteins contain pattern L1, 88 % contain pattern M2, 77 % contain pattern L3 and 2 % contain pattern M4.

Family **H** (Bacterial CopA Proteins) contains proteins of which 98 % were of bacterial origin (homologous families H1 - H6). 83 % are annotated as laccases in GenBank. 50 % contain pattern L1, 50 % M2, 42 % L3 and 3 % M4.

Family **I** (Bilirubin Oxidases) contains proteins of which 70 % were of bacterial origin (homologous families I1 - I5), 15 % of plant origin (homologous family I3), 10 % of fungal origin (homologous family I1) and 5 % of unspecified source organism (figure 6.5.2). 3 % are annotated in GenBank as laccases. 70 % contain pattern L1, 92 % M2, 91 % L3 and 65 % M4.

Family **J** (Bacterial CueO Proteins) contains proteins of which 90 % were of bacterial origin (homologous families J1 - J4) and 10 % of eukaryotic origin (homologous family J5). 12 % are annotated as laccases in GenBank. 74 % of the proteins contain pattern L1, 75 % M2, 76 % L3 and 3 % M4.

Besides the slight variations within the laccase and MCO sequence patterns almost all MCO sequences share the same highly conserved copper binding residues (figure 6.5.1). They could be identified and annotated by a manual validation of each family-specific multisequence alignment. Only within homologous families F5, F6, H3 and J2 these residues could not be detected.

Discussion

As suggested previously (Hoegger et al., 2006), the ten MCO superfamilies were named by combining the name of the prevailing source organism and the putative enzymatic function. The overall distribution of source organisms among the families generally agreed with the initial classification (figure 6.5.2). Since the assignment of a protein to a superfamily was exclusively based on sequence similarity, it was expected to find in the same family proteins from different source organisms, even from different kingdoms of life. Indeed, most families consisted of a majority of proteins belonging to one kingdom with a minority from other kingdoms, despite a sequence similarity as high as 40 %. A systematic classification of all proteins by sequence similarity only was prerequisite to a reliable sequence alignment of superfamilies and to identify conserved, functionally relevant sequence patterns.

However, a systematic analysis of previously described MCO- and laccase-specific patterns (Kumar et al., 2003; Messerschmidt and Huber, 1990; Ouzounis and Sander, 1991), which have been derived from a small number of MCOs and laccases, demonstrated their low sensitivity (table 6.5.3). The MCO patterns M2 and M4 were only found in 9 and 65 % of all MCOs (table 6.5.2), respectively. 8 % of all MCOs contain both M2 and M4. To differentiate laccases from other MCOs is even more difficult. If we assume that sequence similarity is an indication of function similarities, there are four superfamilies which contain putative laccases. However, for these families the laccase-specific patterns L1 and L3 were only found in 45 % and 37 % of the sequences, respectively (table 6.5.2). Only 8 % of all putative laccases contain all four patterns simultaneously. This low percentage of positive hits could either indicate that "laccase superfamilies" contain MCOs without laccases activity, or it might be caused by the too restrictive patterns. As an alternative to patterns, sequence profiles are widely used to specify functionally related protein families (Servant et al., 2002; Finn et al., 2008). Therefore, for each superfamily a hidden Markov profile is provided, and the four copper binding regions are consistently annotated in the *LccED*.

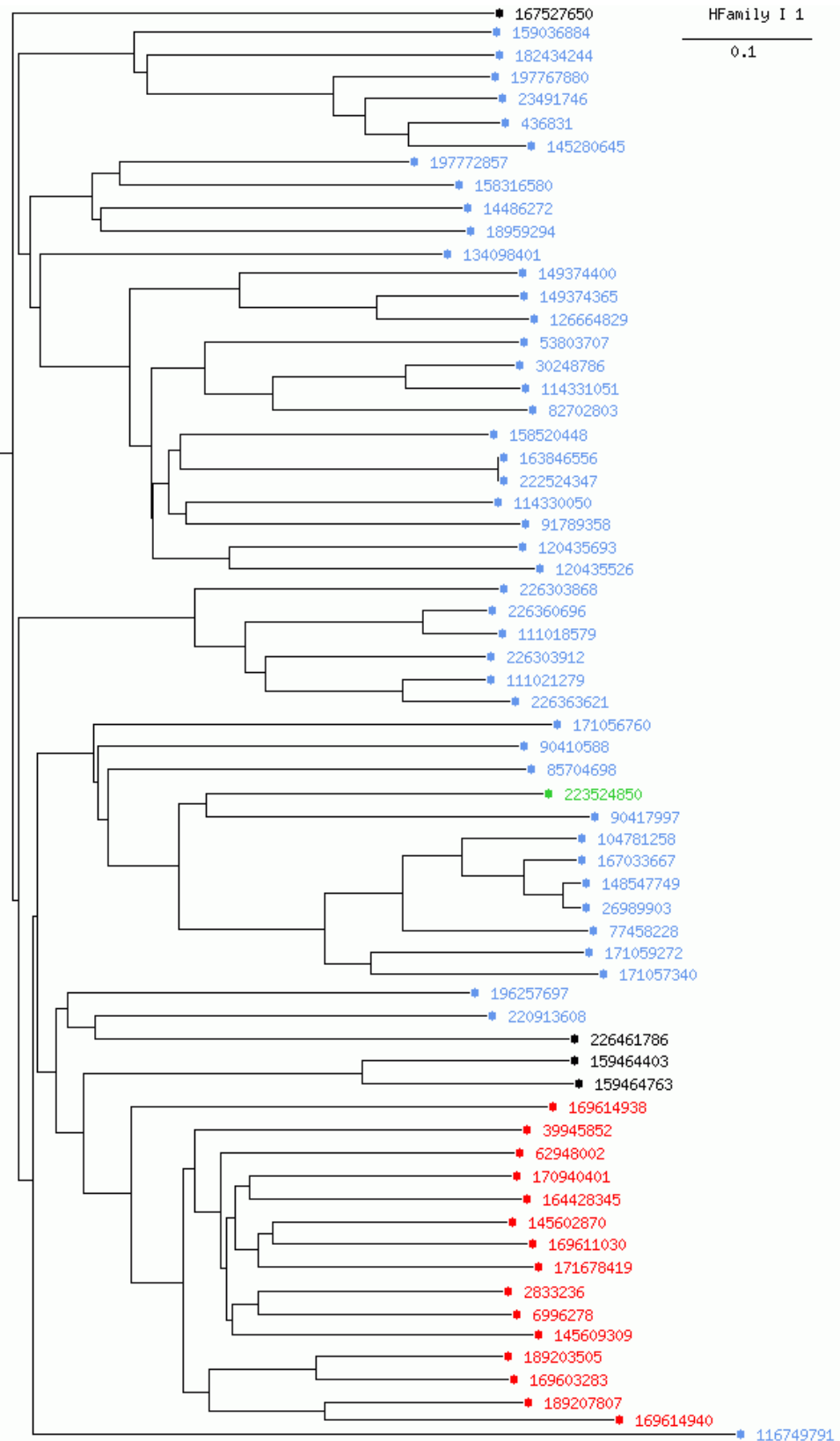


Figure 6.5.2: Phylogenetic tree for the homologous family I1 (Bilirubin Oxidases). The chosen coloring option is "by kingdom". Entries of bacterial origin are shown in blue, fungal entries in red, plant proteins in green and non-specified entries are colored in black.

To define discriminating rules for laccases, more detailed functional studies are needed. It has been shown previously that a systematic classification of large protein families based on sequence similarity and comprehensive analysis tools as provided by the *LccED* serve as a reliable framework for studying sequence-structure-function relationships of enzyme families (Fischer and Pleiss, 2003; Knoll et al., 2009; Sirim et al., 2009) and for the design of mutants or focused mutant libraries with improved biochemical properties (Seifert and Pleiss, 2008; Seifert et al., 2009).

Table 6.5.2: Counted patterns L1 - L4 for each superfamily

Superfamily	Proteins	L1	M2	L3	M4
A (Basidiomycete Laccases)	201	146	149	166	95
B (Ascomycete Laccases)	421	37	20	198	329
C (Insect Laccases)	168	94	12	26	140
D (Fungal Pigment MCOs)	55	–	11	5	46
E (Fungal Ferroxidases)	117	1	62	83	98
F (Fungal and plant AOs)	137	21	2	47	63
G (Plant Laccases)	333	232	–	27	257
H (Bacterial CopA Proteins)	383	–	96	29	173
I (Bacterial Bilirubin Oxidases)	149	–	–	–	35
J (Bacterial CueO Proteins)	310	–	–	–	233

Table 6.5.3: Active site regions missed by the patterns referred as false negatives

Superfamily	Proteins	L1	M2	L3	M4
A (Basidiomycete Laccases)	201	46	44	29	93
B (Ascomycete Laccases)	421	371	385	205	65
C (Insect Laccases)	168	50	127	127	5
D (Fungal Pigment MCOs)	55	50	43	45	6
E (Fungal Ferroxidases)	117	97	46	27	4
F (Fungal and plant AOs)	137	121	77	90	10
G (Plant Laccases)	333	49	292	256	6
H (Bacterial CopA Proteins)	383	190	193	158	13
I (Bacterial Bilirubin Oxidases)	149	105	137	136	97
J (Bacterial CueO Proteins)	310	231	233	235	10

6.5.6 Conclusion

The Laccase Engineering Database enables the systematic classification and analysis of MCO sequences and structures from different public sources. The integration of protein data in a relational database system has been used to study the molecular basis of biochemical properties and to investigate sequence-structure-function relationships. The *LccED* comes with a set of tools for phylogenetic analysis and classification. The annotated multisequence alignments allow the identification of the regions which house the copper atoms and other functionally relevant residues.

6.5.7 Availability and requirements

The *LccED* is available at <http://www.LccED.uni-stuttgart.de>. Via this web interface all sequences, alignments and trees are accessible and all data is supplied for download.

6.5.8 Authors' contributions

DS carried out the study, the establishment and the analysis of the *LccED*, and wrote the manuscript. FW extended and executed the *DWARF* applications, and generated the web interface. LW performed the analysis and contributed to the manuscript. RDS contributed to the discussion. JP supervised the project and finalized the manuscript. All authors read and approved the final version of the manuscript.

6.5.9 Acknowledgements

This work was supported by the German Research Foundation (SFB706).

6.5.10 References

- Alexandre, G., Zhulin, I. B., Laccases are widespread in bacteria. *Trends Biotechnol* 18 (2), 41–42, 2000.
- Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25 (17), 3389–3402, 1997.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L., GenBank. *Nucleic Acids Res* 36 (Database issue), D25–D30, 2008.

- Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., Zardecki, C., The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 58 (Pt 6 No 1), 899–907, 2002.
- Bourbonnais, R., Paice, M. G., Oxidation of non-phenolic substrates. An expanded role for laccase in lignin biodegradation. *FEBS Lett* 267 (1), 99–102, 1990.
- Clutterbuck, A. J., Absence of laccase from yellow-spored mutants of *Aspergillus nidulans*. *J Gen Microbiol* 70 (3), 423–435, 1972.
- Couto, S. R., Herrera, J. L. T., Industrial and biotechnological applications of laccases: a review. *Biotechnol Adv* 24 (5), 500–513, 2006.
- Delano, W. L., The PyMOL Molecular Graphics System. San Carlos, CA, USA: DeLano Scientific, 2002.
- Dittmer, N. T., Suderman, R. J., Jiang, H., Zhu, Y.-C., Gorman, M. J., Kramer, K. J., Kanost, M. R., Characterization of cDNAs encoding putative laccase-like multicopper oxidases and developmental expression in the tobacco hornworm, *Manduca sexta*, and the malaria mosquito, *Anopheles gambiae*. *Insect Biochem Mol Biol* 34 (1), 29–41, 2004.
- Durbin, R., Eddy, S., Krogh, A., Mitchison, G., Biological sequence analysis. Cambridge University Press, 1998.
- Finn, R. D., Tate, J., Mistry, J., Coghill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., Bateman, A., The Pfam protein families database. *Nucleic Acids Res* 36 (Database issue), D281–D288, 2008.
- Fischer, M., Pleiss, J., The Lipase Engineering Database: a navigation and analysis tool for protein families. *Nucleic Acids Res* 31 (1), 319–321, 2003.
- Fischer, M., Thai, Q. K., Grieb, M., Pleiss, J., DWARF—a data warehouse system for analyzing protein families. *BMC Bioinformatics* 7, 495, 2006.
- Geiger, J. P., Nicole, M., Nandris, D., Rio, B., Root-Rot Diseases of *Hevea-Brasiliensis*. 1. Physiological and Biochemical Aspects of Host Aggression. *European Journal of Forest Pathology* 16, 22–37, 1986.

- Hoegger, P. J., Kilaru, S., James, T. Y., Thacker, J. R., Kuees, U., Phylogenetic comparison and classification of laccase and related multicopper oxidase protein sequences. *FEBS J* 273 (10), 2308–2326, 2006.
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuče, B. A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P. S., Sigrist, C. J. A., The 20 years of PROSITE. *Nucleic Acids Res* 36 (Database issue), D245–D249, 2008.
- Kabsch, W., Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12), 2577–2637, 1983.
- Knoll, M., Hamm, T. M., Wagner, F., Martinez, V., Pleiss, J., The PHA Depolymerase Engineering Database: A systematic analysis tool for the diverse family of polyhydroxyalkanoate (PHA) depolymerases. *BMC Bioinformatics* 10, 89, 2009.
- Kumar, S. V. S., Phale, P. S., Durani, S., Wangikar, P. P., Combined sequence and structure analysis of the fungal laccase family. *Biotechnol Bioeng* 83 (4), 386–394, 2003.
- Kunamneni, A., Camarero, S., García-Burgos, C., Plou, F. J., Ballesteros, A., Alcalde, M., Engineering and Applications of fungal laccases for organic synthesis. *Microb Cell Fact* 7, 32, 2008.
- Mayer, A. M., Staples, R. C., Laccase: new functions for an old enzyme. *Phytochemistry* 60 (6), 551–565, 2002.
- Messerschmidt, A. (Ed.), *Multi-Copper Oxidases*. World Scientific Pub Co Inc, 1997.
- Messerschmidt, A., Huber, R., The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin. Modelling and structural relationships. *Eur J Biochem* 187 (2), 341–352, 1990.
- Murphy, M. E., Lindley, P. F., Adman, E. T., Structural comparison of cupredoxin domains: domain recycling to construct proteins with novel functions. *Protein Sci* 6 (4), 761–770, 1997.
- Nakamura, K., Go, N., Function and molecular evolution of multicopper blue proteins. *Cell Mol Life Sci* 62 (18), 2050–2066, 2005.
- O'Malley, D. M., Whetten, R., Bao, W., Chen, C.-L., Sederoff, R. R., The role of laccase in lignification. *Plant Journal* 4, 751 – 757, 1993.

- Ouzounis, C., Sander, C., A structure-derived sequence pattern for the detection of type I copper binding domains in distantly related proteins. *FEBS Lett* 279 (1), 73–78, 1991.
- Piontek, K., Antorini, M., Choinowski, T., Crystal structure of a laccase from the fungus *Trametes versicolor* at 1.90-Å resolution containing a full complement of coppers. *J Biol Chem* 277 (40), 37663–37669, 2002.
- Rice, P., Longden, I., Bleasby, A., EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16 (6), 276–277, 2000.
- Riva, S., Laccases: blue enzymes for green chemistry. *Trends Biotechnol* 24 (5), 219–226, 2006.
- Seifert, A., Pleiss, J., Identification of selectivity-determining residues in cytochrome P450 monooxygenases: A systematic analysis of the substrate recognition site 5. *Proteins*, 2008.
- Seifert, A., Vomund, S., Grohmann, K., Kriening, S., Urlacher, V. B., Laschat, S., Pleiss, J., Rational design of a minimal and highly enriched CYP102A1 mutant library with improved regio-, stereo- and chemoselectivity. *Chembiochem* 10 (5), 853–861, 2009.
- Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D., Kahn, D., ProDom: automated clustering of homologous domains. *Brief Bioinform* 3 (3), 246–251, 2002.
- Sharma, P., Goel, R., Capalash, N., Bacterial laccases. *World J Microbiol Biotechnol* 23, 823–832, 2007.
- Sigrist, C. J. A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P., PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 3 (3), 265–274, 2002.
- Sirim, D., Wagner, F., Lisitsa, A., Pleiss, J., The cytochrome P450 engineering database: Integration of biochemical properties. *BMC Biochem* 10, 27, 2009.
- Solomon, E., Sundaram, U., Machonkin, T., Multicopper Oxidases and Oxygenases. *Chem Rev* 96 (7), 2563–2606, 1996.
- Thompson, J. D., Higgins, D. G., Gibson, T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22 (22), 4673–4680, 1994.

7 Gesamtliteraturverzeichnis

- Adman, E. T., Copper protein structures. *Adv Protein Chem* 42, 145–197, 1991.
- Alexandre, G., Zhulin, I. B., Laccases are widespread in bacteria. *Trends Biotechnol* 18 (2), 41–42, 2000.
- Alexov, E. G., Gunner, M. R., Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys J* 72 (5), 2075–2093, 1997.
- Altschul, S. F., Madden, T. L., Schaeffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25 (17), 3389–3402, 1997.
- Appel, D., Schmid, R. D., Dragan, C. A., Bureik, M., Urlacher, V. B., A fluorimetric assay for cortisol. *Anal. Bioanal. Chem.* 383 (2), 182–6, 2005.
- Apweiler, R., Bairoch, A., Wu, C. H., Protein sequence databases. *Curr Opin Chem Biol* 8 (1), 76–80, 2004.
- Baudry, J., Rupasinghe, S., Schuler, M. A., Class-dependent sequence alignment strategy improves the structural and functional modeling of P450s. *Protein Eng Des Sel* 19 (8), 345–353, 2006.
- Beck, J. S., Vartuli, J. C., Roth, W. J., Leonowicz, M. E., Kresge, C. T., Schmitt, K. D., Chu, C. T. W., Olson, D. H., Sheppard, E. W., Mccullen, S. B., Higgins, J. B., Schlenker, J. L., A New Family of Mesoporous Molecular-Sieves Prepared with Liquid-Crystal Templates. *J. Am. Chem. Soc.* 114 (27), 10834–10843, 1992.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L., GenBank. *Nucleic Acids Res* 31 (1), 23–27, 2003.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L., GenBank. *Nucleic Acids Res* 36 (Database issue), D25–D30, 2008.

- Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* 28, 235–242, 2000.
- Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., Zardecki, C., The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 58 (Pt 6 No 1), 899–907, 2002.
- Bernhardt, R., Cytochrome P450: structure, function, and generation of reactive oxygen species. *Rev Physiol Biochem Pharmacol* 127, 137–221, 1996.
- Bernhardt, R., Cytochromes P450 as versatile biocatalysts. *J Biotechnol* 124 (1), 128–145, 2006.
- Bertrand, T., Jolival, C., Briozzo, P., Caminade, E., Joly, N., Madzak, C., Mougin, C., Crystal structure of a four-copper laccase complexed with an arylamine: insights into substrate recognition and correlation with kinetics. *Biochemistry* 41 (23), 7325–7333, 2002.
- Betts, M. J., Russell, R. B., In: *Bioinformatics for Geneticists*. M. R. Barnes and I. C. Gray, Wiley, 2003.
- Boger, T., Roesky, R., Glaser, R., Ernst, S., Eigenberger, G., Weitkamp, J., Influence of the aluminum content on the adsorptive properties of MCM-41. *Microporous Mat.* 8 (1-2), 79–91, 1997.
- Bourbonnais, R., Paice, M. G., Oxidation of non-phenolic substrates. An expanded role for laccase in lignin biodegradation. *FEBS Lett* 267 (1), 99–102, 1990.
- Burton, S. G., Oxidizing enzymes as biocatalysts. *Trends Biotechnol* 21 (12), 543–549, 2003.
- Case, D. A., Darden, T. A., T.E. Cheatham, I., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Pearlman, D. A., Crowley, M., Walker, R. C., Zhang, W., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Wong, K. F., Paesani, F., Wu, X., Brozell, S., Tsui, V., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Beroza, P., Mathews, D. H., Schafmeister, C., Ross, W. S., Kollman, P. A., Amber 9. University of California, 2006.

- Chalupský, J., Neese, F., Solomon, E. I., Ryde, U., Rulísek, L., Multireference ab initio calculations on reaction intermediates of the multicopper oxidases. *Inorg Chem* 45 (26), 11051–11059, 2006.
- Chiang, C., Yeh, H., Wang, L., Chan, N., Crystal Structure of the Human Prostacyclin Synthase. *J. Mol. Biol.* 364, 266–274, 2006.
- Choi, M., Heo, W., Kleitz, F., Ryoo, R., Facile synthesis of high quality mesoporous SBA-15 with enhanced control of the porous network connectivity and wall thickness. *Chem. Commun. (Camb)* (12), 1340–1, 2003.
- Chothia, C., Lesk, A. M., The relation between the divergence of sequence and structure in proteins. *EMBO J* 5 (4), 823–826, 1986.
- Cirino, P., Arnold, F., Regioselectivity and activity of cytochrome P450 BM-3 and mutant F87A in reactions driven by hydrogen peroxide. *Adv. Synth. Catal.* 344, 932–937, 2002.
- Cirino, P. C., Arnold, F. H., A Self-Sufficient Peroxide-Driven Hydroxylation Biocatalyst. *Angew. Chem. Int. Ed. Engl.* 42 (28), 3299–3301, 2003.
- Clutterbuck, A. J., Absence of laccase from yellow-spored mutants of *Aspergillus nidulans*. *J Gen Microbiol* 70 (3), 423–435, 1972.
- Codd, E. F., Data models in database management. *Proceedings of the 1980 Workshop on Data Abstraction, Databases and Conceptual Modeling* 11, 112–114, 1980.
- Couto, S. R., Herrera, J. L. T., Industrial and biotechnological applications of laccases: a review. *Biotechnol Adv* 24 (5), 500–513, 2006.
- Cramer, C., Truhlar, D., Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem Rev* 99 (8), 2161–2200, 1999.
- de Graaf, C., Vermeulen, N. P. E., Feenstra, K. A., Cytochrome p450 in silico: an integrative modeling approach. *J Med Chem* 48 (8), 2725–2755, 2005.
- Deere, J., Magner, E., Wall, J. G., Hodnett, B. K., Mechanistic and structural features of protein adsorption onto mesoporous silicates. *J. Phys. Chem. B* 106 (29), 7340–7347, 2002.
- Delano, W. L., *The PyMOL Molecular Graphics System*. San Carlos, CA, USA: DeLano Scientific, 2002.

- Denisov, I. G., Makris, T. M., Sligar, S. G., Schlichting, I., Structure and chemistry of cytochrome P450. *Chem Rev* 105 (6), 2253–2277, 2005.
- Diaz, J. F., Balkus, K. J., Enzyme immobilization in MCM-41 molecular sieve. *J. Mol. Catal. B: Enzym.* 2 (2-3), 115–126, 1996.
- Dittmer, N. T., Suderman, R. J., Jiang, H., Zhu, Y.-C., Gorman, M. J., Kramer, K. J., Kanost, M. R., Characterization of cDNAs encoding putative laccase-like multicopper oxidases and developmental expression in the tobacco hornworm, *Manduca sexta*, and the malaria mosquito, *Anopheles gambiae*. *Insect Biochem Mol Biol* 34 (1), 29–41, 2004.
- Doolittle, R. F., Protein sequence comparisons: searching databases and aligning sequences. *Curr Opin Biotechnol* 5 (1), 24–28, 1994.
- Durbin, R., Eddy, S., Krogh, A., Mitchison, G., Biological sequence analysis. Cambridge University Press, 1998.
- Eddy, S. R., Profile hidden Markov models. *Bioinformatics* 14 (9), 755–763, 1998.
- Eddy, S. R., A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol* 4 (5), e1000069, 2008.
- Elmasri, E. R., Navathe, S., Fundamentals of Database Systems. Addison Wesley; 5 edition, 2006.
- Farinas, E. T., Schwaneberg, U., Glieder, A., Arnold, F. H., Directed evolution of a cytochrome P450 monooxygenase for alkane oxidation. *Adv. Synth. Catal.* 343 (6-7), 601–606, 2001.
- Feng, D. F., Doolittle, R. F., Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* 25 (4), 351–360, 1987.
- Finn, R. D., Tate, J., Mistry, J., Coghill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., Bateman, A., The Pfam protein families database. *Nucleic Acids Res* 36 (Database issue), D281–D288, 2008.
- Fischer, M., Knoll, M., Sirim, D., Wagner, F., Funke, S., Pleiss, J., The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. *Bioinformatics* 23 (15), 2015–2017, 2007.

- Fischer, M., Pleiss, J., The Lipase Engineering Database: a navigation and analysis tool for protein families. *Nucleic Acids Res* 31 (1), 319–321, 2003.
- Fischer, M., Thai, Q. K., Grieb, M., Pleiss, J., DWARF—a data warehouse system for analyzing protein families. *BMC Bioinformatics* 7, 495, 2006.
- Gaasterland, T., Structural genomics taking shape. *Trends Genet* 14 (4), 135, 1998.
- Garfinkel, D., Studies on pig liver microsomes. I. Enzymic and pigment composition of different microsomal fractions. *Arch Biochem Biophys* 77 (2), 493–509, 1958.
- Geiger, J. P., Nicole, M., Nandris, D., Rio, B., Root-Rot Diseases of Hevea-Brasiliensis .1. Physiological and Biochemical Aspects of Host Aggression. *European Journal of Forest Pathology* 16, 22–37, 1986.
- Georgescu, R. E., Alexov, E. G., Gunner, M. R., Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. *Biophys J* 83 (4), 1731–1748, 2002.
- Giammona, D. A., Case, D., Bayly, C., Force field modifications for all-atom heme. University of California, 1984.
- Girhard, M., Schuster, S., Dietrich, M., Dürre, P., Urlacher, V. B., Cytochrome P450 monooxygenase from *Clostridium acetobutylicum*: a new alpha-fatty acid hydroxylase. *Biochem Biophys Res Commun* 362 (1), 114–119, 2007.
- Gotoh, O., Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J Biol Chem* 267 (1), 83–90, 1992.
- Gotoh, O., Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol* 264 (4), 823–838, 1996.
- Graham, S. E., Peterson, J. A., How similar are P450s and what can their differences teach us? *Arch Biochem Biophys* 369 (1), 24–29, 1999.
- Guengerich, F. P., Reactions and Significance of Cytochrome P-450 Enzymes. *Journal of Biological Chemistry* 266 (16), 10019 – 10022, 1991.

- Guengerich, F. P., Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem Res Toxicol* 14 (6), 611–650, 2001.
- Guengerich, F. P., Rate-limiting steps in cytochrome P450 catalysis. *Biol Chem* 383 (10), 1553–1564, 2002.
- Hannemann, F., Bichet, A., Ewen, K. M., Bernhardt, R., Cytochrome P450 systems—biological variations of electron transport chains. *Biochim Biophys Acta* 1770 (3), 330–344, 2007.
- Hasemann, C. A., Kurumbail, R. G., Boddupalli, S. S., Peterson, J. A., Deisenhofer, J., Structure and function of cytochromes P450: a comparative analysis of three crystal structures. *Structure* 3 (1), 41–62, 1995.
- Hernández, M. C. R., Wejebe, J. E. M., Alcáñtara, J. I. V., Ruvalcaba, R. M., Serrano, L. A. G. a., Ferrara, J. T., Immobilization of cytochrome P-450 on MCM-41 with different silicon/aluminum ratios. *Microporous and Mesoporous Materials* 80, 25–31, 2005.
- Ho, W. W., Li, H., Nishida, C. R., de Montellano, P. R. O., Poulos, T. L., Crystal structure and properties of CYP231A2 from the thermoacidophilic archaeon *Picrophilus torridus*. *Biochemistry* 47 (7), 2071–2079, 2008.
- Hoegger, P. J., Kilaru, S., James, T. Y., Thacker, J. R., Kuees, U., Phylogenetic comparison and classification of laccase and related multicopper oxidase protein sequences. *FEBS J* 273 (10), 2308–2326, 2006.
- Hudson, S., Cooney, J., Magner, E., Proteins in mesoporous silicates. *Angew. Chem. Int. Ed. Engl.* 47 (45), 8582–94, 2008.
- Hudson, S., Magner, E., Cooney, J., Hodnett, B. K., Methodology for the immobilization of enzymes onto mesoporous materials. *J Phys Chem B* 109 (41), 19496–19506, 2005.
- Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P. S., Sigrist, C. J. A., The 20 years of PROSITE. *Nucleic Acids Res* 36 (Database issue), D245–D249, 2008.
- Humphrey, W., Dalke, A., Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* 14 (1), 33–8, 27–8, 1996.

- Kabsch, W., Sander, C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22 (12), 2577–2637, 1983.
- Karplus, K., Barrett, C., Hughey, R., Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14 (10), 846–856, 1998.
- Katiyar, A., Yadav, S., Smirniotis, P. G., Pinto, N. G., Synthesis of ordered large pore SBA-15 spherical particles for adsorption of biomolecules. *J. Chromatogr. A* 1122 (1-2), 13–20, 2006.
- Kemper, B., Structural basis for the role in protein folding of conserved proline-rich regions in cytochromes P450. *Toxicol Appl Pharmacol* 199 (3), 305–315, 2004.
- Khan, K. K., Mazumdar, S., Modi, S., Sutcliffe, M., Roberts, G. C., Mitra, S., Steady-state and picosecond-time-resolved fluorescence studies on the recombinant heme domain of *Bacillus megaterium* cytochrome P-450. *Eur. J. Biochem.* 244 (2), 361–70, 1997.
- King, D. L., Azari, M. R., Wiseman, A., Immobilization of cytochrome P-450 enzyme from *Saccharomyces cerevisiae*. *Methods Enzymol.* 137, 675–686, 1988.
- Klingenberg, M., Pigments of rat liver microsomes. *Arch Biochem Biophys* 75 (2), 376–386, 1958.
- Knoll, M., Hamm, T. M., Wagner, F., Martinez, V., Pleiss, J., The PHA Depolymerase Engineering Database: A systematic analysis tool for the diverse family of polyhydroxyalkanoate (PHA) depolymerases. *BMC Bioinformatics* 10, 89, 2009.
- Knoll, M., Pleiss, J., The Medium-Chain Dehydrogenase/reductase Engineering Database: a systematic analysis of a diverse protein family to understand sequence-structure-function relationship. *Protein Sci* 17 (10), 1689–1697, 2008.
- Kresge, C. T., Leonowicz, M. E., Roth, W. J., Vartuli, J. C., Beck, J. S., Ordered Mesoporous Molecular-Sieves Synthesized by a Liquid-Crystal Template Mechanism. *Nature* 359 (6397), 710–712, 1992.
- Krogh, A., Two methods for improving performance of an HMM and their application for gene finding. *Proc Int Conf Intell Syst Mol Biol* 5, 179–186, 1997.
- Kruk, M., Jaroniec, M., Ko, C. H., Ryoo, R., Characterization of the porous structure of SBA-15. *Chem. Mater.* 12 (7), 1961–1968, 2000.

- Kumar, S. V. S., Phale, P. S., Durani, S., Wangikar, P. P., Combined sequence and structure analysis of the fungal laccase family. *Biotechnol Bioeng* 83 (4), 386–394, 2003.
- Kunamneni, A., Camarero, S., García-Burgos, C., Plou, F. J., Ballesteros, A., Alcalde, M., Engineering and Applications of fungal laccases for organic synthesis. *Microb Cell Fact* 7, 32, 2008.
- Lamb, S. B., Lamb, D. C., Kelly, S. L., Stuckey, D. C., Cytochrome P450 immobilisation as a route to bioremediation/biocatalysis. *FEBS Lett.* 431 (3), 343–6, 1998.
- Landwehr, M., Hochrein, L., Otey, C. R., Kasrayan, A., Backvall, J. E., Arnold, F. H., Enantioselective alpha-hydroxylation of 2-arylacetic acid derivatives and buspirone catalyzed by engineered cytochrome P450 BM-3. *J. Am. Chem. Soc.* 128 (18), 6058–6059, 2006.
- Larrondo, L. F., Salas, L., Melo, F., Vicuna, R., Cullen, D., A novel extracellular multicopper oxidase from *Phanerochaete chrysosporium* with ferroxidase activity. *Appl Environ Microbiol* 69 (10), 6257–6263, 2003.
- Lawton, T. J., Sayavedra-Soto, L. A., Arp, D. J., Rosenzweig, A. C., Crystal structure of a two-domain multicopper oxidase: implications for the evolution of multicopper blue proteins. *J Biol Chem* 284 (15), 10174–10180, 2009.
- Lee, C. H., Lang, J., Yen, C. W., Shih, P. C., Lin, T. S., Mou, C. Y., Enhancing stability and oxidation activity of cytochrome c by immobilization in the nanochannels of mesoporous aluminosilicates. *J. Phys. Chem. B.* 109 (25), 12277–12286, 2005.
- Lee, D.-S., Yamada, A., Sugimoto, H., Matsunaga, I., Ogura, H., Ichihara, K., Adachi, S.-I., Park, S.-Y., Shiro, Y., Substrate recognition and molecular mechanism of fatty acid hydroxylation by cytochrome P450 from *Bacillus subtilis*. Crystallographic, spectroscopic, and mutational studies. *J Biol Chem* 278 (11), 9761–9767, 2003.
- Li, H., Poulos, T. L., The structure of the cytochrome p450BM-3 haem domain complexed with the fatty acid substrate, palmitoleic acid. *Nat Struct Biol* 4 (2), 140–146, 1997.
- Li, H. M., Mei, L. H., Urlacher, V. B., Schmid, R. D., Cytochrome P450 BM-3 evolved by random and saturation mutagenesis as an effective indole-hydroxylating catalyst. *Appl Biochem Biotechnol* 144 (1), 27–36, 2008.

- Li, Q. S., Ogawa, J., Schmid, R. D., Shimizu, S., Engineering cytochrome P450 BM-3 for oxidation of polycyclic aromatic hydrocarbons. *Appl. Environ. Microbiol.* 67 (12), 5735–9., 2001a.
- Li, Q. S., Ogawa, J., Shimizu, S., Critical role of the residue size at position 87 in H₂O₂-dependent substrate hydroxylation activity and H₂O₂ inactivation of cytochrome P450BM-3. *Biochem. Biophys. Res. Commun.* 280 (5), 1258–1261, 2001b.
- Lisitsa, A. V., Gusev, S. A., Karuzina, I. I., Archakov, A. I., Koymans, L., Cytochrome P450 database. *SAR QSAR Environ Res* 12 (4), 359–366, 2001.
- Mast, N., Graham, S. E., Andersson, U., Bjorkhem, I., Hill, C., Peterson, J., Pikuleva, I. A., Cholesterol binding to cytochrome P450 7A1, a key enzyme in bile acid biosynthesis. *Biochemistry* 44 (9), 3259–3271, 2005.
- Maurer, S., Urlacher, V., Schulze, H., Schmid, R. D., Immobilisation of P450 BM-3 and an NADP⁺ cofactor recycling system: towards a technical application of heme-containing monooxygenases in fine chemical synthesis. *Adv. Synth. Catal.* 345, 802–810, 2003.
- May, A. C., Optimal classification of protein sequences and selection of representative sets from multiple alignments: application to homologous families and lessons for structural genomics. *Protein Eng* 14 (4), 209–217, 2001.
- Mayer, A. M., Staples, R. C., Laccase: new functions for an old enzyme. *Phytochemistry* 60 (6), 551–565, 2002.
- McLean, K. J., Sabri, M., Marshall, K. R., Lawson, R. J., Lewis, D. G., Clift, D., Balding, P. R., Dunford, A. J., Warman, A. J., McVey, J. P., Quinn, A. M., Sutcliffe, M. J., Scrutton, N. S., Munro, A. W., Biodiversity of cytochrome P450 redox systems. *Biochem Soc Trans* 33 (Pt 4), 796–801, 2005.
- Messerschmidt, A. (Ed.), 1997. *Multi-Copper Oxidases*. World Scientific Pub Co Inc.
- Messerschmidt, A., Huber, R., The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin. Modelling and structural relationships. *Eur J Biochem* 187 (2), 341–352, 1990.
- Mestres, J., Structure conservation in cytochromes P450. *Proteins* 58 (3), 596–609, 2005.
- Miyahara, M., Vinu, A., Ariga, K., Immobilization of lysozyme onto pore-engineered mesoporous AISBA-15. *J Nanosci Nanotechnol* 6 (6), 1765–1771, 2006.

- Montellano, O. d., Cytochrome P450: structure, mechanism and biochemistry. New York, Plenum Press, 1995.
- Munro, A. W., Girvan, H. M., McLean, K. J., Cytochrome P450-redox partner fusion enzymes. *Biochim Biophys Acta* 1770 (3), 345–359, 2007.
- Munro, A. W., Leys, D. G., McLean, K. J., Marshall, K. R., Ost, T. W. B., Daff, S., Miles, C. S., Chapman, S. K., Lysek, D. A., Moser, C. C., Page, C. C., Dutton, P. L., P450 BM3: the very model of a modern flavocytochrome. *Trends Biochem Sci* 27 (5), 250–257, 2002.
- Murphy, M. E., Lindley, P. F., Adman, E. T., Structural comparison of cupredoxin domains: domain recycling to construct proteins with novel functions. *Protein Sci* 6 (4), 761–770, 1997.
- Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., Venter, J. C., A whole-genome assembly of *Drosophila*. *Science* 287 (5461), 2196–2204, 2000.
- Nakamura, K., Go, N., Function and molecular evolution of multicopper blue proteins. *Cell Mol Life Sci* 62 (18), 2050–2066, 2005.
- Nebert, D. W., Adesnik, M., Coon, M. J., Estabrook, R. W., Gonzalez, F. J., Guengerich, F. P., Gunsalus, I. C., Johnson, E. F., Kemper, B., Levin, W., The P450 gene superfamily: recommended nomenclature. *DNA* 6 (1), 1–11, 1987.
- Nebert, D. W., Nelson, D. R., Adesnik, M., Coon, M. J., Estabrook, R. W., Gonzalez, F. J., Guengerich, F. P., Gunsalus, I. C., Johnson, E. F., Kemper, B., The P450 superfamily: updated listing of all genes and recommended nomenclature for the chromosomal loci. *DNA* 8 (1), 1–13, 1989.
- Needleman, S. B., Wunsch, C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48 (3), 443–453, 1970.
- Nelson, D. R., Mining databases for cytochrome P450 genes. *Methods Enzymol* 357, 3–15, 2002.

- Nelson, D. R., Cytochrome P450 nomenclature, 2004. *Methods Mol Biol.* 320, 1–10, 2006.
- Nicholls, A., Honig, B., A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J Comp Chem* 12, 435–445, 1991.
- Noh, H., Yohe, S. T., Vogler, E. A., Volumetric interpretation of protein adsorption: ion-exchange adsorbent capacity, protein pI, and interaction energetics. *Biomaterials* 29 (13), 2033–2048, 2008.
- O'Connor, A. J., Hokura, A., Jenny M. Kisler, S. S., Stevens, G. W., Komatsu, Y., Amino acid adsorption onto mesoporous silica molecular sieves. *Separation and Purification Technology* 48, 197–201, 2006.
- O'Malley, D. M., Whetten, R., Bao, W., Chen, C.-L., Sederoff, R. R., The role of laccase in lignification. *Plant Journal* 4, 751 – 757, 1993.
- Omura, T., Sato, R., The carbon monoxide-binding pigment of liver microsomes. *J. Biol. Chem.* 239 (7), 2370–2378, 1964.
- Oscarson, M., Ingelman-Sundberg, M., CYPalleles: a web page for nomenclature of human cytochrome P450 alleles. *Drug Metab Pharmacokinet* 17 (6), 491–495, 2002.
- O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., Notredame, C., 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol* 340 (2), 385–395, 2004.
- Ouzounis, C., Sander, C., A structure-derived sequence pattern for the detection of type I copper binding domains in distantly related proteins. *FEBS Lett* 279 (1), 73–78, 1991.
- Park, J., Lee, S., Choi, J., Ahn, K., Park, B., Park, J., Kang, S., Lee, Y.-H., Fungal cytochrome P450 database. *BMC Genomics* 9, 402, 2008.
- Peterson, J. A., Graham, S. E., A close family resemblance: the importance of structure in understanding cytochromes P450. *Structure* 6 (9), 1079–1085, 1998.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., Ferrin, T. E., UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25 (13), 1605–1612, 2004.

- Picard, R. R., Cook, R. D., Cross-Validation of Regression Models. *Journal of the American Statistical Association* 79 (387), 575–583, 1984.
- Piontek, K., Antorini, M., Choinowski, T., Crystal structure of a laccase from the fungus *Trametes versicolor* at 1.90-Å resolution containing a full complement of coppers. *J Biol Chem* 277 (40), 37663–37669, 2002.
- Pires, E. L., Miranda, E. A., Valenca, G. P., Gas-phase enzymatic esterification on immobilized lipases in MCM-41 molecular sieves. *Appl. Biochem. Biotech.* 98, 963–976, 2002.
- Poulos, T. L., Finzel, B. C., Howard, A. J., High-resolution crystal structure of cytochrome P450cam. *J Mol Biol* 195 (3), 687–700, 1987.
- Raucy, J. L., Allen, S. W., Recent advances in P450 research. *Pharmacogenomics J* 1 (3), 178–186, 2001.
- Rice, P., Longden, I., Bleasby, A., EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16 (6), 276–277, 2000.
- Riva, S., Laccases: blue enzymes for green chemistry. *Trends Biotechnol* 24 (5), 219–226, 2006.
- Rosales-Hernández, M., Kispert, L., Torres-Ramírez, E., Ramírez-Rosales, D., Zamorano-Ulloa, R., Trujillo-Ferrara, J., Electron paramagnetic resonance analyses of biotransformation reactions with cytochrome P-450 immobilized on mesoporous molecular sieves. *Biotechnol Lett* 29 (6), 919–924, 2007.
- Rosenzweig, A. C., Sazinsky, M. H., Structural insights into dioxygen-activating copper enzymes. *Curr Opin Struct Biol* 16 (6), 729–735, 2006.
- Rossmann, M. G., Argos, P., Exploring structural homology of proteins. *J Mol Biol* 105 (1), 75–95, 1976.
- Rupasinghe, S., Schuler, M. A., Kagawa, N., Yuan, H., Lei, L., Zhao, B., Kelly, S. L., Waterman, M. R., Lamb, D. C., 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 (27), 6338–6342, 2006.

- Russell, R. B., Barton, G. J., Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins* 14 (2), 309–323, 1992.
- Sali, A., Blundell, T. L., Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234 (3), 779–815, 1993.
- Salis, A., Meloni, D., Ligas, S., Casula, M. F., Monduzzi, M., Solinas, V., Dumitriu, E., Physical and chemical adsorption of *Mucor javanicus* lipase on SBA-15 mesoporous silica. Synthesis, structural characterization, and activity performance. *Langmuir* 21 (12), 5511–5516, 2005.
- Sayari, A., Kruk, M., Jaroniec, M., Characterization of microporous-mesoporous MCM-41 silicates prepared in the presence of octyltrimethylammonium bromide. *Catal. Lett.* 49 (3-4), 147–153, 1997.
- Scheeff, E. D., Bourne, P. E., Application of protein structure alignments to iterated hidden Markov model protocols for structure prediction. *BMC Bioinformatics* 7, 410, 2006.
- Schmid, A., Dordick, J. S., Hauer, B., Kiener, A., Wubbolts, M., Witholt, B., Industrial biocatalysis today and tomorrow. *Nature* 409 (6817), 258–268, 2001.
- Schwaneberg, U., Schmidt-Dannert, C., Schmitt, J., Schmid, R. D., A continuous spectrophotometric assay for P450 BM-3, a fatty acid hydroxylating enzyme, and its mutant F87A. *Anal. Biochem.* 269 (2), 359–66, 1999.
- Schwede, T., Diemand, A., Guex, N., Peitsch, M. C., Protein structure computing in the genomic era. *Res Microbiol* 151 (2), 107–112, 2000.
- Seifert, A., Pleiss, J., Identification of selectivity-determining residues in cytochrome P450 monooxygenases: A systematic analysis of the substrate recognition site 5. *Proteins*, 2008.
- Seifert, A., Vomund, S., Grohmann, K., Kriening, S., Urlacher, V. B., Laschat, S., Pleiss, J., Rational design of a minimal and highly enriched CYP102A1 mutant library with improved regio-, stereo- and chemoselectivity. *ChemBiochem* 10 (5), 853–861, 2009.
- Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D., Kahn, D., ProDom: automated clustering of homologous domains. *Brief Bioinform* 3 (3), 246–251, 2002.

- Sevrioukova, I. F., Li, H., Zhang, H., Peterson, J. A., Poulos, T. L., Structure of a cytochrome P450-redox partner electron-transfer complex. *Proc Natl Acad Sci U S A* 96 (5), 1863–1868, 1999.
- Sharma, P., Goel, R., Capalash, N., Bacterial laccases. *World J Microbiol Biotechn* 23, 823–832, 2007.
- Sigrist, C. J. A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P., PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 3 (3), 265–274, 2002.
- Sirim, D., Wagner, F., Lisitsa, A., Pleiss, J., The cytochrome P450 engineering database: Integration of biochemical properties. *BMC Biochem* 10, 27, 2009.
- Sitkoff, D., Sharp, K. A., Honig, B., Accurate calculation of hydration free energies using macroscopic solvent models. *J Phys Chem* 98, 1978–1988, 1994.
- Smith, T. F., Waterman, M. S., Identification of common molecular subsequences. *J Mol Biol* 147 (1), 195–197, 1981.
- Solomon, E., Sundaram, U., Machonkin, T., Multicopper Oxidases and Oxygenases. *Chem Rev* 96 (7), 2563–2606, 1996.
- Solomon, E. I., Chen, P., Metz, M., Lee, S.-K., Palmer, A. E., Oxygen Binding, Activation, and Reduction to Water by Copper Proteins. *Angew Chem Int Ed Engl* 40 (24), 4570–4590, 2001.
- Song, S. W., Hidajat, K., Kawi, S., Functionalized SBA-15 materials as carriers for controlled drug delivery: Influence of surface properties on matrix-drug interactions. *Langmuir* 21 (21), 9568–9575, 2005.
- Strushkevich, N., Tempel, W., Dombrowski, L., Dong, A., Loppnau, P., Arrowsmith, C., Edwards, A., Bountra, C., Wilkstrom, M., Bochkarev, A., Park, H., Crystal structure of human CYP7A1. To be Published, PDB: 2DAX.
- Tadesse, M. A., D'Annibale, A., Galli, C., Gentili, P., Sergi, F., An assessment of the relative contributions of redox and steric issues to laccase specificity towards putative substrates. *Org Biomol Chem* 6 (5), 868–878, 2008.

- Takahashi, H., Li, B., Sasaki, T., Miyazaki, C., Kajino, T., Inagaki, S., Catalytic activity in organic solvents and stability of immobilized enzymes depend on the pore size and surface characteristics of mesoporous silica. *Chem. Mater.* 12 (11), 3301–3305, 2000.
- Takahashi, H., Li, B., Sasaki, T., Miyazaki, C., Kajino, T., Inagaki, S., Immobilized enzymes in ordered mesoporous silica materials and improvement of their stability and catalytic activity in an organic solvent. *Microporous Mesoporous Mater.* 44, 755–762, 2001.
- Taylor, M., Lamb, D. C., Cannell, R. J., Dawson, M. J., Kelly, S. L., Cofactor recycling with immobilized heterologous cytochrome P450 105D1 (CYP105D1). *Biochem. Biophys. Res. Commun.* 279 (2), 708–11, 2000.
- Thompson, J. D., Higgins, D. G., Gibson, T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22 (22), 4673–4680, 1994.
- Trodler, P., Nieveler, J., Rusnak, M., Schmid, R. D., Pleiss, J., Rational design of a new one-step purification strategy for *Candida antarctica* lipase B by ion-exchange chromatography. *J Chromatogr A* 1179 (2), 161–167, 2008.
- UniProt Consortium, The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* 37 (Database issue), D169–D174, 2009.
- Urlacher, V., Schmid, R. D., Biotransformations using prokaryotic P450 monooxygenases. *Curr Opin Biotechnol* 13 (6), 557–564, 2002.
- Urlacher, V. B., Eiben, S., Cytochrome P450 monooxygenases: perspectives for synthetic application. *Trends Biotechnol* 24, 324–330, 2006.
- Urlacher, V. B., Makhsomkhanov, A., Schmid, R. D., Biotransformation of beta-ionone by engineered cytochrome P450 BM-3. *Appl Microbiol Biotechnol* 70 (1), 53–59, 2006.
- Venter, J. C., Adams, M. D., Myers, E. W., The sequence of the human genome. *Science* 291 (5507), 1304–1351, 2001.
- Vinu, A., Murugesan, V., Hartmann, M., Adsorption of lysozyme over mesoporous molecular sieves MCM-41 and SBA-15: Influence of pH and aluminum incorporation. *J. Phys. Chem. B* 108 (22), 7323–7330, 2004a.

- Vinu, A., Murugesan, V., Tangermann, O., Hartmann, M., Adsorption of cytochrome c on mesoporous molecular sieves: Influence of pH, pore diameter, and aluminum incorporation. *Chem. Mater.* 16 (16), 3056–3065, 2004b.
- Vinu, A., Streb, C., Murugesan, V., Hartmann, M., Adsorption of cytochrome c on new mesoporous carbon molecular sieves. *J. Phys. Chem. B* 107 (33), 8297–8299, 2003.
- Wade, R. C., Motiejunas, D., Schleinkofer, K., Sudarko, Winn, P. J., Banerjee, A., Kariakin, A., Jung, C., Multiple molecular recognition mechanisms. Cytochrome P450—a case study. *Biochim Biophys Acta* 1754 (1-2), 239–244, 2005.
- Walker, R. C., Crowley, M. F., Case, D. A., The implementation of a fast and accurate QM/MM potential method in Amber. *J Comput Chem* 29 (7), 1019–1031, 2008.
- Wang, J., Cieplak, P., Kollman, P. A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* 21, 1049–1074, 2000.
- Werck-Reichhart, D., Feyereisen, R., Cytochromes P450: a success story. *Genome Biol* 1 (6), REVIEWS3003, 2000.
- Widmann, M., Clairo, M., Dippon, J., Pleiss, J., Analysis of the distribution of functionally relevant rare codons. *BMC Genomics* 9, 207, 2008.
- Yoon, B. J., Lenhoff, A. M., Computation of the electrostatic interaction energy between a protein and a charged surface. *J Phys Chem* 96, 3130–3134, 1992.
- Yoshida, H., Chemistry of lacquer (urushi). *Journal of the Chemical Society* 43, 472–486, 1883.
- Zuckerandl, E., Pauling, L., Molecules as documents of evolutionary history. *J Theor Biol* 8 (2), 357–366, 1965.