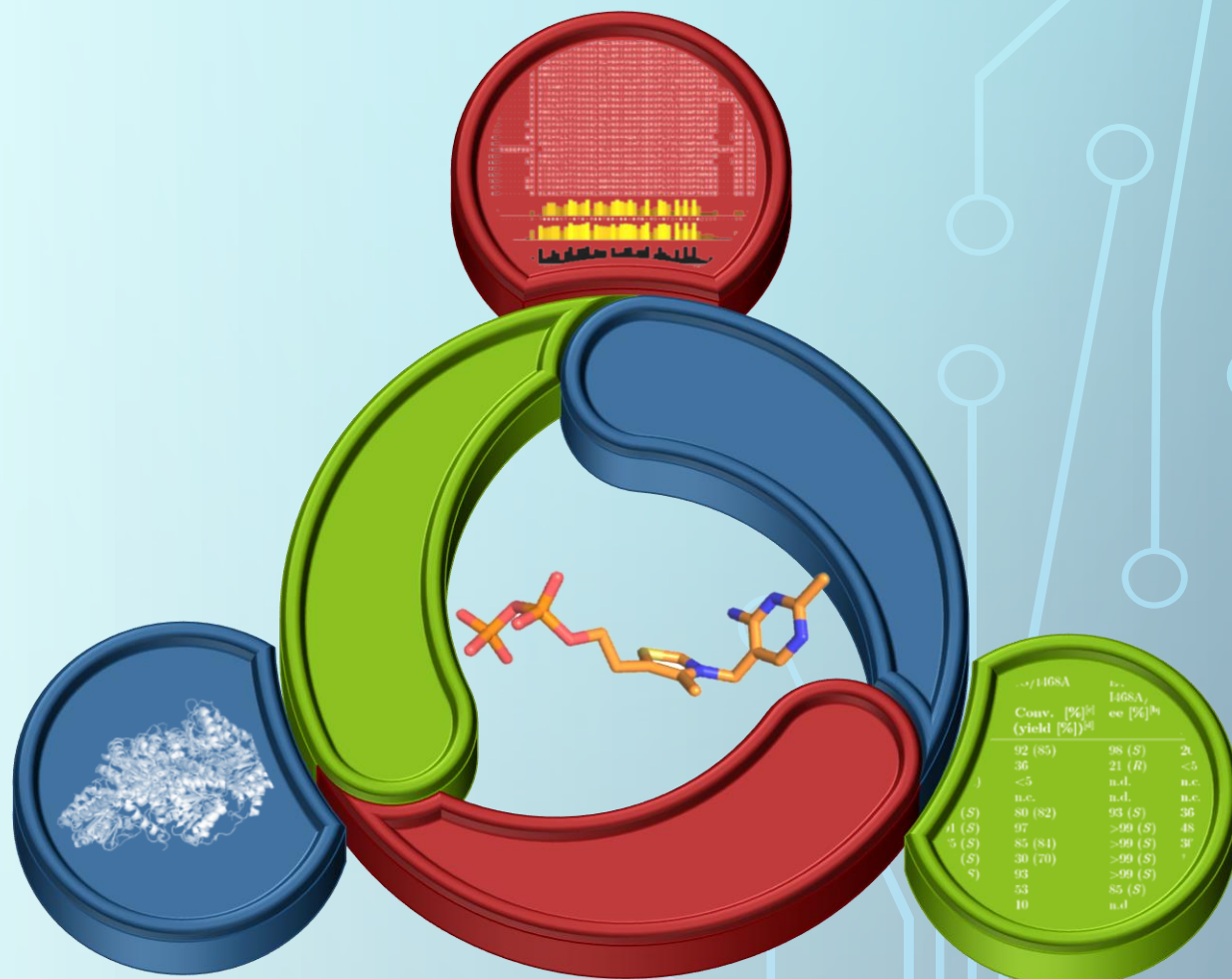


Constantin Vogel

Systematic Analysis of the Sequence-Structure-Function Relationships of Thiamine Diphosphate-dependent Enzymes



Systematic Analysis of the Sequence-Structure-Function Relationships of Thiamine Diphosphate-dependent Enzymes

Systematische Analyse der Sequenz-Struktur-Funktions Zusammenhänge bei
Thiamindiphosphat-abhängigen Enzymen

Von der Fakultät 4: Energie-, Verfahrens- und Biotechnik der Universität Stuttgart
zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung.

Vorgelegt von
Constantin Vogel
aus Bad Reichenhall

Hauptberichter: Prof. Dr. Jürgen Pleiss
Mitberichter: Prof. Dr.-Ing. Ralf Takors
Prüfungsvorsitz: Prof. Dr. Bernhard Hauer

Tag der mündlichen Prüfung: 27. Februar 2015

Institut für Technische Biochemie der Universität Stuttgart

2015

The cover picture contains elements provided by others than the author of this work:

- The circle arrow was adopted from the '3-D cyclical shapes' template provided by Microsoft Office within the 'Free professionally-designed templates for PowerPoint 2010' package.
- The structure of thiamine diphosphate was taken from the crystal structure of the pyruvate decarboxylase from *Acetobacter pasteurianus* (ApPDC, pdb|2VBI, Rother et al. 2011) and visualized in PyMol (The PyMOL Molecular Graphics System, Version 1.7.0.5 Schrödinger, LLC.).
- The depicted alignment was generated from branched-chain α -ketoacid decarboxylase sequences taken from the Thiamine diphosphate-dependent Enzyme Engineering Database (TEED) (Widmann, Radloff, and Pleiss 2010; Vogel and Pleiss 2014) using *clustalo* (Sievers et al. 2011) and was visualized in JalView (Waterhouse et al. 2009).
- The superimposed structures of ApPDC and the benzaldehyde lyase from *Pseudomonas fluorescens* (pdb|2AG0, Mosbacher, Müller, and Schulz 2005) were aligned and visualized using PyMol (The PyMOL Molecular Graphics System, Version 1.7.0.5 Schrödinger, LLC.).
- The table showing results of the experimental characterization of ApPDC-variants was taken from the publication Westphal et al. 2014a, which is part of this work (Table 4.5 on page 125).

David Lipman about Margaret Dayhoff (* March 11, 1925; † February 5, 1983):

"She established the three major components of what a bioinformaticist does: a mixture of their own basic discoveries with the data, which are biological discoveries; tool development, where they share those tools with other people; and resource development. She did all three, and she did incredibly important things in all three."

Moody, Glyn (2004). *Digital code of life: How bioinformatics is revolutionizing science, medicine, and business*. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, p. 11.

Acknowledgements

This thesis and the practical and theoretical work behind it was generously supported by many people. I would like to express my sincere gratitude to those people, who directly became part of my PhD projects, provided and were part of the pleasant working environment and accompanied me for the last years.

I would like to express my sincere gratitude to my supervisor Prof. Dr. Jürgen Pleiss for the opportunity to work on this project, for teaching and support in scientific writing, for the inspiring, motivating discussions, for helpful suggestions, for providing the needed computational infrastructure and for his receptiveness for new ideas. Moreover, I would like to thank for the freedom I had to develop own ideas, to work on different projects in different collaborations and for the trust in my person.

I would like to thank Prof. Dr. Bernhard Hauer for the opportunity to do my doctorate at the Institute of Technical Biochemistry. The institute provided an inspiring environment and encouraged discussions on various topics, which allowed to think out of the box. Furthermore, I thank Prof. Dr. Bernhard Hauer for his willingness to be the chair of examiners in the defense of this thesis.

I would like to thank Prof. Dr.-Ing. Ralf Takors from the Institute of Biochemical Engineering for the kind effort to be co-referee of this thesis.

Furthermore, I would like to express my sincere gratitude to Prof. Dr. Martina Pohl from the Institute of Bio- and Geosciences, IBG-1, Department of Biocatalysis and Biosensors, at the Forschungszentrum Jülich GmbH, for the critical discussions and helpful suggestions that scrutinized various ideas and smoothed the resulting projects. I further thank for the support while screening the available literature about variants of ThDP-dependent enzymes. Finally yet importantly, I sincerely appreciate the motivating recognition I received.

I would also like to extend my appreciation to all the members of the DFG Research unit 1296 "Diversity of Asymmetric Thiamine Catalysis" for the familiar and fruitful environment, which enabled beneficial cooperation. Especially Prof. Dr. Michael Müller, Dr. Dörte Rother and Prof. Dr. Kai Tittmann for the critical but constructive and pleasant discussions. In this context, I would also like to thank the DFG (Deutsche Forschungsgemeinschaft) for funding of this Research Unit.

Special thanks to my (former) PhD colleagues from the DFG Research Unit 1296 Dr. Robert Westphal, Dr. Sabrina Loschonsky, Dr. Cindy Wechsler, Alexander Fries, Thomas Broja, Shiromi Baier, Kai Mahnken, Saskia Bock, Anna Baierl, Simon Waltzer, Lydia Walter, Rüdiger Ohs, Fabian von Pappenheim and Dr. Maryam Beigi for the familiar atmosphere, the fruitful collaborations and for the PhD meetings, which I will keep in good memory.

I would particularly like to thank my friend Robert for his extensive support, the constant motivation and his willingness to cooperate. Bioinformatics is help- and powerful and did for good reason develop to a distinct field in science. However, without experimental application of findings gained from bioinformatics, its practical utility often might remain vague. Thus, I sincerely appreciate Robert's efforts to incorporate my ideas into experimental biocatalysis.

I would also like to express my gratitude to Saskia Bock and Anna Baierl, who both likewise started ongoing cooperation.

My sincere appreciation is further extended to my colleagues in the Bioinformatics working group at the Institute of Technical Biochemistry at the University of Stuttgart. Sincere thanks to Dr. Michael Widmann, for fruitful discussions and collaboration on standard numbering schemes and integration of biochemical data into family-specific protein databases. I would also like to thank Silvia Fademrecht for countless discussions, lots of motivation and her contribution to the *BioCatNet* system by intensive testing and reporting of bugs. Special thanks for their help with molecular dynamics (MD) simulations to Tobias Kulschewski, who introduced me into the world of MD, Sven Benson, who provided the force field of the ThDP cofactor, and Łukasz Gricman. Especially I would like to express my gratitude to Łukasz Gricman for the cooperation on the database of p450 monooxygenases, the numbering scheme for this protein family and his efforts on the development of a literature mining tool, which helped me to compile information about variants of ThDP-dependent enzymes. I would also like to thank Florian Wagner for a brief introduction into the DWARF system and his work on the parsing and pushing algorithms, which were partially reused in the algorithms developed during the last years. Furthermore, I

would like to thank all other (including former) members of the Bioinformatics working group for the willingness to share knowledge, programs and documentations that enabled constructive work in science as well as in management of the IT infrastructure.

I further express my sincere acknowledgments to the students I was permitted to supervise and whose results interlace this thesis. Thanks to Chantal Göttler for her work on *SERgrid*, Lenz Lorenz for his work on the development of tools for the identification of modularity in transaldolases, Hannah Dienhart for her work on *DBUpdate* and Waldemar Reusch for playing a key role in the development of the *BioCatNet* system. Many people mentioned on those pages deserve sincere gratitude for their involvement in the development of the relational data model and the definition of the minimal requirements, but I especially thank Waldemar for his outstanding dedication to this project, the excellent work on the graphical user interface, his participation in configuration and management of the web server and his constructive contribution to the development of *BioCatNet*.

I would also like to thank Yahayda Aladzeme and Sebastian Enderle, who contributed to the analysis of the modular structure of ThDP-dependent enzymes during an internship. Further, I would like to thank my former research assistants Jelena Ochs, Essam Abdelhady and Catharina Zeil for their work on the programming of web interfaces, which had a beneficial effect on the development of the *BioCatNet* system. I would also like to extend my appreciation to all the members of the Institute of Technical Biochemistry for the inspiring discussions, and Christine Klumpp-Klug for the support in organizational issues.

Furthermore, I would like to thank Silvia, Ewelina and Robert for thorough proof-reading of this thesis.

Finally, I would like to express my sincere gratitude to my family and friends, especially to my wife Ewelina, who accompanied me during my studies and my doctorate and were unstinting with support, sympathy, motivation and encouragement whenever it was required.

Contents

Acknowledgements	iv
Contents	vii
Nomenclature	x
List of Figures	xiv
List of Tables	xvii
Abstract/Zusammenfassung	xix
1 Introduction	1
1.1 Sequence - Structure - Function	1
1.2 Online repositories of sequence and structure information	3
1.3 Family-specific protein databases (FSPDs)	5
1.4 Homology modeling	7
1.5 Standard numbering schemes	8
1.6 Thiamine diphosphate-dependent enzymes	9
1.6.1 Thiamine diphosphate	11
1.6.2 Reaction mechanism	13
1.6.3 Regulation of enantio- and chemoselectivity in ThDP-dependent enzymes	13
1.6.4 Sequence and structure of ThDP-dependent enzymes	17
1.6.5 Family classification	18
1.6.6 ThDP-dependent Enzyme Engineering Database (TEED)	19
1.7 Scope and objectives	20

2	Results	22
2.1	Generation and maintenance of family-specific protein databases	22
2.1.1	The Thiamine diphosphate-dependent Enzyme Engineering Database	22
2.1.2	DBParse	23
2.1.3	DBUpdate	27
2.1.4	BioCatNet	28
2.2	Sequence	31
2.2.1	A standard numbering scheme for ThDP-dependent decarboxylases	32
2.2.2	A standard numbering scheme for the PYR and PP domains of ThDP-dependent enzymes	37
2.3	Structure	40
2.3.1	The modular structure of ThDP-dependent enzymes	40
2.3.2	Structural rearrangement of the <i>Ap</i> PDC	41
2.3.3	Automated homology modeling	43
2.4	Function	46
2.4.1	Rational engineering of a ThDP-dependent enzyme for the direct asymmetric synthesis of (<i>S</i>)-benzoins	46
3	Discussion	54
3.1	Computing performance - a big issue in bioinformatics applications	54
3.2	The sequence-structure-function relationships of ThDP-dependent enzymes	57
3.2.1	Sequence-Structure	57
3.2.2	Structure-Function	61
3.2.3	Sequence-Function	64
3.2.4	BioCatNet	72
3.3	Conclusions and future perspectives	74
4	Publications	76
4.1	A standard numbering scheme for thiamine diphosphate-dependent decarboxylases	76
4.1.1	Abstract	76
4.1.2	Background	77
4.1.3	Results	81
4.1.4	Discussion	88

4.1.5	Conclusions	90
4.1.6	Methods	90
4.2	The modular structure of ThDP-dependent enzymes	92
4.2.1	Abstract	92
4.2.2	Introduction	93
4.2.3	Materials and Methods	95
4.2.4	Results	98
4.2.5	Discussion	107
4.2.6	Conclusions	117
4.3	A chimeric ThDP-enzyme for the direct asymmetric synthesis of (<i>S</i>)-benzoins .	119
4.3.1	Abstract	119
4.3.2	Communication	119
4.3.3	Experimental Section	126
4.4	BioCatNet	127
4.4.1	Abstract	127
4.4.2	Background	128
4.4.3	Concept, Implementation and Population	130
4.4.4	Application and future perspectives	142
References		146
A Supporting Information		174
A.1	Structural rearrangement of <i>Ap</i> PDC	174
A.2	SERgrid	175
A.3	Variants of ThDP-dependent enzymes	177
A.4	A standard numbering scheme for thiamine diphosphate-dependent decarboxylases	188
A.5	The modular structure of ThDP-dependent enzymes	210
A.6	A chimeric ThDP-enzyme for the direct asymmetric synthesis of (<i>S</i>)-benzoins .	216
A.7	BioCatNet	225
B List of Publications		229
C Erklärung der eigenständigen Arbeit		232

Nomenclature

2-HPCL	2-Hydroxyphytanoyl-CoA lyases
2-HPP	2-Hydroxypropiophenone
<i>Ap</i>	<i>Acetobacter pasteurianus</i>
<i>As</i>	<i>Azoarcus sp.</i>
<i>At</i>	<i>Agrobacterium tumefaciens</i>
<i>Bp</i>	<i>Burkholderia pseudomallei</i>
<i>Ec</i>	<i>Escherichia coli</i>
<i>ee</i>	Enantiomeric excess
<i>Ll</i>	<i>Lactococcus lactis</i>
<i>Pf</i>	<i>Pseudomonas fluorescens</i>
<i>Pp</i>	<i>Pseudomonas putida</i>
<i>Sc</i>	<i>Saccharomyces cerevisiae</i>
<i>Td</i>	<i>Torulaspora delbrueckii</i>
<i>Zm</i>	<i>Zymomonas mobilis</i>
aa	Amino acid
acc nr	Accession number
ADP	Adenosine diphosphate
AHAS	Acetohydroxyacid synthase
aKADH	α -Ketoacid dehydrogenase
aKGDH	α -Ketoglutarate dehydrogenase
AMP	Adenosine monophosphate
ASIC	Application-specific integrated circuit
ATN	N-terminal acyltransferase-like domain
BAL	Benzaldehyde lyase

BFDC	Benzoylformate decarboxylase
BLAST	Basic local alignment search tool
BLOSUM	Blocks substitution matrix
CDH	Cyclohexane-1,2-dione hydrolase
CDR	Complementary-determining region
CPU	Central processing unit
CYP	Cytochrome P450 monooxygenase
CYPED	CYtochrome P450 Engineering Database
DC	Decarboxylase
DED	PHA depolymerase engineering database
DNA	Deoxyribonucleic acid
DOPE	Discrete optimized protein energy
DWARF	Data warehouse system for analyzing protein families
DXPS	1-Deoxy-D-xylulose-5-phosphate synthase
e.g.	exempli gratia
EC	Enzyme commission
FAD	Flavin adenine dinucleotide
FSPD	Family-specific protein database
gi	GenInfo identifier
GPU	Graphics processing unit
GUI	Graphical user interface
GXC	Glyoxylate carboligase
HMM	Hidden Markov model
HTML	Hypertext markup language
ID	Identifier
IOR	Indolepyruvate:ferredoxin oxidoreductase
IPDC	Indolepyruvate decarboxylase
KGOR	2-Ketoglutarate oxidoreductase
LACED	Lactamase engineering database

LCCED	Laccase and multicopper oxidase engineering database
LED	Lipase engineering database
LIMS	Laboratory information management system
MBLED	Metallo- β -lactamase engineering database
MCO	Multi copper oxidase
MD	Molecular dynamics
MDRED	Medium-chain dehydrogenase/reductase engineering database
MSA	Multisequence alignment
NCBI	National Center for Biotechnology Information
NMR	Nuclear magnetic resonance
nvw	'numbering Vogel-Widmann' file format
OCDC	Oxalyl-CoA decarboxylase
OGDC	2-Oxoglutarate decarboxylase
OGDH	2-Oxoglutarate dehydrogenase
OR	Oxidoreductase
PAC	Phenylacetylcarbinol
PDB	Protein Data Bank
PDC	Pyruvate decarboxylase
PDH	Pyruvate dehydrogenase
PFOR	Pyruvate:ferredoxin oxidoreductase
POX	Pyruvate oxidase
PP	Pyrophosphate-binding domain
PPDC	Phosphonopyruvate decarboxylase
PYR	Pyrimidine-binding domain
RAM	Random-access memory
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RNA	Ribonucleic acid

RSCB	Research Collaboratory for Structural Bioinformatics
SAAT	Sulfoacetaldehyde acetyltransferase
SDS	Sodium dodecyl sulfate
SEPHCHC	2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexadiene-1-carboxylate
SID	Sequence identifier / Sequence ID
sim.	Similarity
sp	SwissProt
SPDC	Sulfopyruvate decarboxylase
STAMP	Structural alignment of multiple proteins
std. pos.	Standard position
TEED	Thiamine diphosphate-dependent enzyme engineering database
TH3	Transhydrogenase dIII domain
ThDP	Thiamine diphosphate
TK	Transketolase
TKC	Transketolase C-terminal domain
TTCED	Triterpene cyclase engineering database
VOR	2-Ketoisovalerate:ferredoxin oxidoreductase
wt	Wild type
XML	Extensible markup language

List of Figures

1.1	The four levels of protein structures	2
1.2	Statistics on the entries in online repositories of sequences and structures of proteins.	4
1.3	Benzaldehyde, acetaldehyde and the chiral products of homo- and crosscoupling	9
1.4	α -hydroxyketones accessible via ThDP-dependent enzymes starting from acetaldehyde and benzaldehyde	10
1.5	Activation of Thiamine diphosphate	12
1.6	Reaction mechanism of ThDP-dependent enzymes	14
1.7	Schematic representation of the donor- and acceptor-binding sites of <i>PfBAL</i> , <i>PpBFD</i> , <i>ZmPDC</i> and <i>ApPDC-E469G</i>	15
1.8	Binding of ThDP in the active site of <i>ApPDC</i>	16
1.9	Monomer and homodimer of <i>ScPDC</i>	18
2.1	Schematic representation of standard number assignment	33
2.2	Order of different domains on the primary structure of ThDP-dependent enzymes	38
2.3	RMSF of an intra-monomer/PP-PYR variant of <i>ApPDC</i>	42
2.4	SDS gel of the crude extract of cells with and without the synthetic gene for a designed <i>ApPDC</i> variant	43
2.5	DOPE scores and RMSD of homology models	45
2.6	Schematic representation of the active site characteristics of <i>ApPDC</i> , <i>PfBAL</i> and a hybrid variant	47

2.7	Structural superimposition of α -helices PP- α E of <i>ApPDC</i> and <i>PfBAL</i>	48
2.8	Sequence alignment of PP- α E of <i>ApPDC</i> -E469G, <i>PfBAL</i> and the hybrid variant	49
2.9	Structural superimposition of <i>ApPDC</i> , <i>ApPDC</i> -E469G, <i>PfBAL</i> , <i>PfBAL</i> -T481G and the models of a hybrid variant	51
3.1	Development of the clock rate of Intel CPUs	55
3.2	Structural equivalence of functionally relevant amino acid side chains with different standard position numbers in two different proteins	62
3.3	Positions mutated in ThDP-dependent enzymes	65
4.1	Standard numbering alignment on the web interface	83
4.2	Analysis of accordance of two multisequence alignments	84
4.3	The five different basic layouts found in the known structures of ThDP-dependent enzymes	100
4.4	Structural architectures found in ThDP-dependent enzymes with available struc- ture information	101
4.5	Naming scheme for the ThDP-binding fold conserved within the PYR and PP domains of ThDP-dependent enzymes	105
4.6	Network representation of the average sequence similarities between the PYR and PP domains of ThDP-dependent enzymes	108
4.7	Relative average sequence similarity of the PYR and PP domains	109
4.8	Proposed evolutionary pathway of ThDP-dependent enzymes	115
4.9	Schematic presentation of the scope and limitations of carboligations catalyzed by <i>ApPDC</i> -E469G and <i>PfBAL</i>	120
4.10	Donor-binding sites of <i>ApPDC</i> -E469G and <i>ApPDC</i> -E469G/T384G	122
4.11	Possible stabilization of the parallel-oriented acceptor benzaldehyde in the active sites of two <i>ApPDC</i> variants	123
4.12	Data model of the <i>BioCatNet</i> system	137

4.13	Schematic representation of the <i>BioCatNet</i> system	138
4.14	<i>Sequence</i> view of the <i>BioCatNet</i> WebGUI	145
A.1	Simplified model of the identification of 'effect points' in enzyme structures. . .	175
A.2	Divergent positions 421 and 422 in the comparison of the alignment methods in <i>ScPDC</i> , <i>AsCDH</i> and <i>ApPDC</i>	188
A.3	'Dissimilar' positions 114 and 115 (HH-motif) in <i>ScPDC</i> , <i>PpBFDC</i> and <i>PfBAL</i>	189
A.4	Influence of pH on the stereoselectivity and conversion of <i>ApPDC</i> -E469G/ T384G/I468A-catalyzed (<i>S</i>)-benzoin formation	222
A.5	Influence of temperature on the stereoselectivity and conversion of <i>ApPDC</i> - E469G/T384G/I468A/W543 catalyzed (<i>S</i>)-benzoin formation	223

List of Tables

2.1	Conserved positions ($\geq 80\%$) in DCs identified using the standard numbering scheme for ThDP-dependent DCs. ^[a]	36
2.2	Amino acid distribution in sequences homologous to <i>ApPDC</i> and <i>PfBAL</i>	52
3.1	Positions predicted to be part of the <i>S</i> -pockets of the respective proteins	68
4.1	Representative proteins used for the standard numbering scheme for ThDP-dependent decarboxylases	81
4.2	Superfamilies of ThDP-dependent enzymes and their functional annotation in the NCBI GenBank	99
4.3	Results of the automatic domain identification in all ThDP-dependent enzymes	102
4.4	Enzymatic synthesis of (<i>S</i>)-benzoin as catalyzed by <i>ApPDC</i> variants.	122
4.5	Synthesis of benzoin as catalyzed by <i>ApPDC</i> variants	125
4.6	Minimal requirements for the submission of biochemical data to <i>BioCatNet</i>	132
4.7	Optional information accepted for submission to <i>BioCatNet</i>	132
4.8	Additional information available for the different hierarchical levels in the <i>Sequence</i> view of the TEED	142
A.1	Variants of ThDP-dependent enzymes identified in the literature.	177
A.2	Variants mutated at positions outside of the PYR and PP domains of ThDP-dependent enzymes identified in the literature	182
A.3	22 positions of interest in ThDP-dependent decarboxylases	190

A.4	Functionally relevant positions in selected ThDP-dependent decarboxylases found in the literature	191
A.5	Seed sequences chosen for the generation of the TEED	210
A.6	Location of the PP and PYR domains in known structures of ThDP-dependent enzymes	213
A.7	Overview over the alteration of secondary structure elements in the PYR and PP domains of ThDP-dependent enzymes	215
A.8	Conservation of aromatic amino acids in ThDP-dependent decarboxylases at positions corresponding to the aromatic cluster observed in <i>ApPDC</i>	219
A.9	Family classification in the TEED	228

Abstract

Thiamine diphosphate (ThDP)-dependent enzymes form a vast and diverse protein family, both in the sequence space and in their functional potential. Of particular interest are the enantioselective C-C bond forming and cleavage reactions catalyzed by those enzymes. In these reaction, different ThDP-dependent enzymes provide distinct enantio- and chemoselectivities with often narrow substrate and product ranges. This specificity, which is beneficial for the enantiopure synthesis of fine chemicals like 2-hydroxy ketones, limits the scope of accessible products. Investigations of crystal structures of different ThDP-dependent decarboxylases revealed steric properties in the active sites of those enzymes to control the enantio- and chemoselectivity (*S*-pocket and donor-acceptor concept). Subsequent application of those concepts by modulation of the steric properties of enzymes' active sites enabled rational engineering of biocatalysts with desired, but often only moderate, non-physiological enantioselectivities.

The major objective of this thesis was to systematically analyze the sequences and structures of this enzyme family and to elucidate the relationships between sequence, structure and function. Detailed understanding of those relationships is pivotal for rational engineering and therefore necessary for the design of biocatalysts with desired selectivities. As compared to the enormous size of this enzyme family only a small number of representatives were experimentally characterized. Even less ThDP-dependent enzymes were modified by mutations in order to analyze effects of distinct amino acid residues and still less were structurally determined.

Since the systematic analysis of the sequence-structure-function relationships requires information on the structure and function of a major fraction of family members, methods were developed and applied to increase the amount of available structure and function information. By making use of homology modeling, putative atom coordinates for enzymes lacking experimentally determined structure information were predicted. In addition, by development of a new database system that combines sequence, structure and function information, the acquisition of

accurate and comparable biochemical data unambiguously linked to the biocatalysts' amino acid sequences was enabled.

Comparability of biochemical data and deduction of functional roles of certain residues requires comparable biochemical data on the one hand and methods to compare residues from different enzymes on the other hand. Introduction of standard numbering schemes for ThDP-dependent enzymes facilitated fast and accurate comparison of structurally equivalent positions without the need for structure information. The findings derived from those analyses accelerated the engineering of enzymes with desired enantio- and chemoselectivities and inter alia enabled the enzymatic, direct asymmetric synthesis of (*S*)-benzoins with excellent *ees*.

Zusammenfassung

Die Familie der Thiamindiphosphat (ThDP)-abhängigen Enzyme ist gleichermaßen sequenziell als auch funktionell vielfältig. Besonderes Interesse wird dieser Familie aufgrund ihrer Fähigkeit zuteil, C-C Bindungs- und Spaltungsreaktionen zu katalysieren. Für einen Einsatz in der Biokatalyse und der Synthese von Feinchemikalien (wie beispielsweise α -Hydroxyketone) zeichnen sie sich zudem durch ihre definierten Substratspektren als auch ihre Enantioselektivität in zahlreichen Reaktionen aus. Allerdings schränken diese Spezifitäten das Spektrum an enzymatisch zugänglichen Produkten ein. Vergleichende Untersuchungen vorhandener Proteinstrukturen verschiedener ThDP-abhängiger Enzyme zeigten Unterschiede in der Form der Substrat-Bindetaschen der unterschiedlichen Vertreter. Die daraus abgeleiteten 'S-pocket'- und 'Donor/Akzeptor'-Konzepte führen diese sterischen Unterschiede und die resultierenden verschiedenen räumlichen Anordnungen der beiden Substrate in Ligationsreaktionen als die Ursache verschiedener Enantio- und Substratpräferenzen an. Auf dieser Grundlage konnten, durch Anpassung der Form der aktiven Taschen, Decarboxylasen mit geänderten Selektivitäten erzeugt werden. Oft allerdings einhergehend mit nur moderaten Stereoselektivitäten in der Katalyse nicht-natürlicher Reaktionen.

Für den Erfolg von Rationalem Design von Biokatalysatoren mit gewünschten Eigenschaften sind detaillierte Kenntnisse über die Sequenz-Struktur-Funktions Zusammenhänge der jeweiligen Proteinfamilie von Bedeutung. Diese Doktorarbeit hatte die systematische Analyse dieser Zusammenhänge in ThDP-abhängigen Enzymen zum Ziel. Eine systematische Analyse von Sequenz-Struktur-Funktions Zusammenhängen erfordert implizit Sequenz-, Struktur- und Funktionsinformation für einen Großteil der zur Familie gehörenden Enzyme. In bisherigen Arbeiten wurden - relativ zu den enormen Ausmaßen dieser Proteinfamilie - nur wenige Vertreter experimentell charakterisiert. Für weiterführende Untersuchungen bezüglich des Einflusses bestimmter Aminosäure-Positionen auf die katalytische Aktivität oder Selektivität wurden nochmals nur wenige dieser Enzyme herangezogen. Eine experimentelle Bestimmung der Proteinstruktur,

welche für Rationales Design von Biokatalysatoren von besonderer Bedeutung ist, wurde nur für einen noch geringeren Bruchteil der ThDP-abhängigen Enzyme durchgeführt.

Um dem bestehenden Mangel an Informationen über die Struktur und Funktion von Enzymen zu begegnen, wurden im Rahmen dieser Arbeit Proteinstrukturen per Homologie-Modellierung vorhergesagt und Methoden zur Erfassung und Auswertung von Funktionsdaten entwickelt. Mit Hilfe eines neuartigen Datenbank-Systems zur Erfassung verlässlicher und vergleichbarer Daten über die Funktion und Sequenz von Enzymen, wurde die Basis für eine systematische Analyse der genannten Zusammenhänge geschaffen.

Neben der Verfügbarkeit von Funktionsinformation, eindeutig mit der Sequenz des entsprechenden Enzyms verknüpft, erfordert die systematische Analyse möglicher funktioneller Bedeutungen einzelner Aminosäure-Positionen eine Methode zum Vergleich von Aminosäuren aus verschiedenen Enzymen. Eine solche Methode wurde mit dem hier präsentierten '*standard numbering scheme*' (Standard-Nummerierungs System) zur Verfügung gestellt. Die Anwendung dieser Methode erlaubt die schnelle und akkurate Identifikation strukturell äquivalenter Positionen in verschiedenen Enzymen ohne Abhängigkeit von Strukturinformation zu den jeweils analysierten Proteinen. Die aus diesen Analysen gezogenen Erkenntnisse wurden eingesetzt, um Biokatalysatoren mit gewünschten Enantio- und Chemoselektivitäten zu erzeugen und erstmals die enzymatische, direkte asymmetrische Synthese von (*S*)-Benzoinen zu ermöglichen.

1 Introduction

1.1 Sequence - Structure - Function

In 1952, Kaj Ulrik Linderstrøm-Lang proposed to classify observed structural elements of proteins by the use of four hierarchical levels (Figure 1.1 on the following page) (Linderstrøm-Lang 1952):

- Primary structure: The sequence of covalently coupled amino acids forming a polypeptide.
- Secondary structure: A three-dimensional structure formed by hydrogen bond interactions between carbonyl and imide groups of the primary structure. The most frequent observed secondary structure elements are coils, turns, α -helices and β -sheets formed by β -strands.
- Tertiary structure: By interactions of amino acid side-chains, multiple secondary structure elements of a polypeptide chain fold into a more complex three-dimensional structure.
- Quaternary structure: Multimerization of several identical or different tertiary structures into homo- or heteromultimers, respectively.

This classification inherently describes a relationship between the sequence and the structure of a protein by calling the pure amino acid sequence the 'primary structure'. Amino acid sequences of proteins automatically assemble three-dimensional structures in aqueous solution by interactions of backbone and side-chain atoms of different residues. Thus, the amino acid sequence determines the structural conformation of a protein (Anfinsen 1973). Moreover, similar sequences were observed to also fold into similar structures (Chothia and Lesk 1986; Rost 1999), with only a small number of exceptions (Kosloff and Kolodny 2008). However, similar structures are not necessarily encoded by similar sequences. Although the estimations of the number of different folds vary, it is generally agreed that it is limited (Chothia 1992; Orengo, Jones, and Thornton

1994; Russell 2002). Consequently, even proteins that vary considerably in their sequences can fold into similar structures. In this context, a distinction is made between homologous structures found in proteins with similar sequences caused by a common evolutionary lineage and analogous structures of proteins without detectable evolutionary relationship (Fitch 1970).

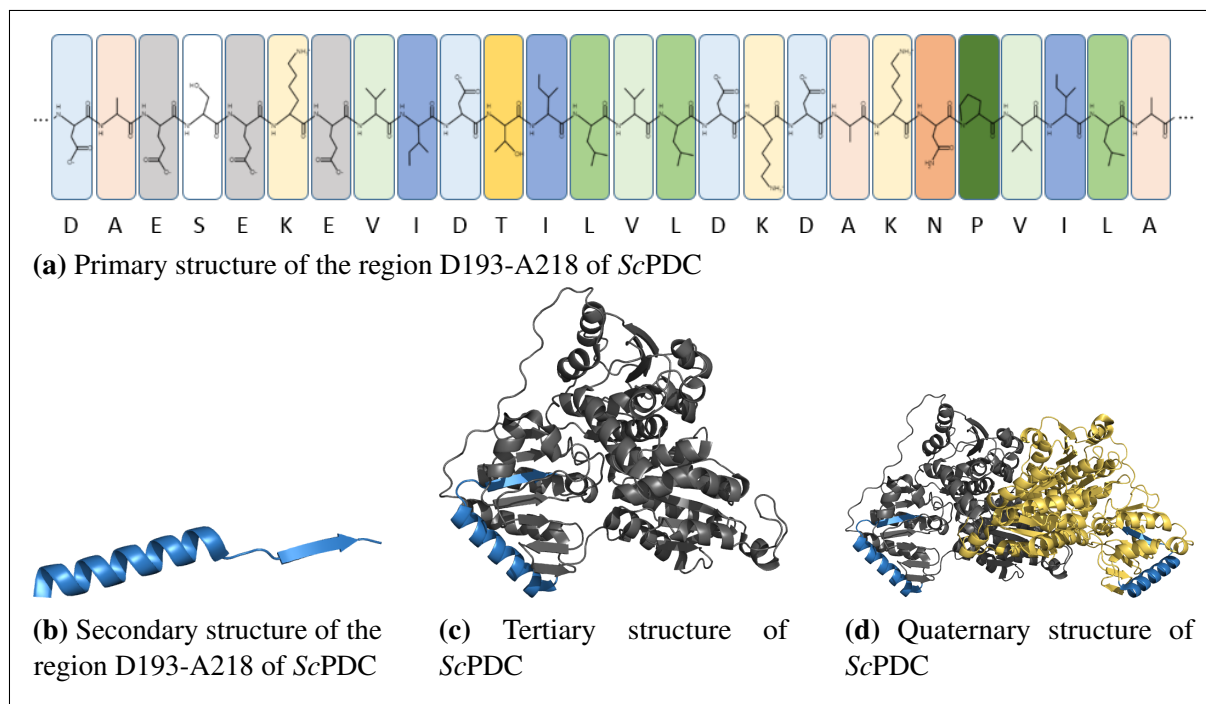


Figure 1.1: The four levels of protein structures defined by Kaj Ulrik Linderstrøm-Lang (Linderstrøm-Lang 1952). (a) The primary structure of a protein is the polypeptide formed by single amino acids that are coupled by peptide bonds. Exemplary, the primary structure of region D193-A218 of the pyruvate decarboxylase from *Saccharomyces cerevisiae* (ScPDC, pdb|2VK8, Kutter et al. 2009) is shown. The polypeptide of the region D193-A218 of ScPDC was drawn using PepDraw (<http://www.tulane.edu/~biochem/WW/PepDraw/>). (b) In aqueous solution, the primary structure folds into the secondary structure. Region D193-A218 of ScPDC consists of an α -helix and a β -strand linked by a short loop. (c) The tertiary structure of ScPDC is the three-dimensional structure of one monomer (gray). Region D193-A218 (blue) folds into a more complex structure in which the β -strand P214-A218 forms part of a β -sheet. (d) By assembly of sequence-identical monomers to a homomultimer, the quaternary structure of ScPDC is formed. In ScPDC, only the quaternary structure possesses catalytic activity.

The catalytic function of enzymes, seen as nano-sized machines, is influenced by their sequence and their structure. Assembled from individual parts, the amino acids, and folded into a three-dimensional structure, the resulting enzyme obtains functionality. Although lowering of the activation energy in the conversion of substrates to products by stabilization of transition states is the basis for catalysis in all enzymes, the catalytic activity of enzymes arises from different factors. Enzymes can provide a catalytic machinery via interaction of several amino acids with specific physicochemical properties or they obtain catalytic activity by the recruitment of cofactors. Examples of amino acid-based reactions are the hydrolysis reactions carried out by lipases (Schrag et al. 1991) and epoxide hydrolases (Barth et al. 2004b). Examples for

proteins catalyzing reactions just by stabilization of energetically unfavored transition states are catalytic antibodies (Lerner, Benkovic, and Schultz 1991), which neither require catalytic amino acids nor cofactors. The reaction occurs spontaneously after binding of the substrate to the epitope. In contrast, enzymes like Cytochrome P450 monooxygenases (Meunier, Visser, and Shaik 2004) and Thiamine diphosphate (ThDP)-dependent enzymes (Frank, Leeper, and Luisi 2007) recruit cofactors like metal ions and/or more complex molecules supporting the physiological reaction mechanism. However, in all cases, the three-dimensional structure confers activity by arranging amino acids and cofactors at specific positions. Moreover, by orchestrating a multitude of amino acids, enzymes form complex channels and active site pockets steering chemo- and enantioselectivity, activity and allow for regulation. However, the complex principles that determine how proteins fold from their sequences into the physiological structures have not been fully elucidated yet. Thus, understanding of enzyme's functions and the vision of designing those nanomachines from scratch requires structure information of proteins with known function.

1.2 Online repositories of sequence and structure information

Since experimental characterization of protein structures by X-ray crystallography or NMR is more challenging and costly than the determination of the nucleotide sequences of DNA or RNA and thus the amino acid sequence of proteins, there is a discrepancy in the numbers of known sequences and structures. In August 2014, there were around 173 million nucleotide and 47 million different amino acid sequences¹ deposited in the GenBank (Benson et al. 2011) of the National Center for Biotechnology Information (NCBI), around 70 million amino acid sequences provided by the UniProtKB/TrEMBL protein database (The UniProt Consortium 2014) and around 100,000 crystal structures from about 66,000 different sequences available at the RCSB Protein Data Bank (PDB) (Berman et al. 2000) (Figure 1.2 on the next page). NCBI and UniProt as well as the PDB provide access to the available information via application programming interfaces (API), which allow for simple, high-performance access and download. All information is provided in various file formats including XML (Extensible Markup Language), which is well-suited since XML parsing modules are available for most if not all

¹The number of different amino acid sequences was counted from the non-redundant BLAST database file provided by the National Center for Biotechnology Information (NCBI) at <ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/>.

programming languages. Entries in those online repositories, regardless of whether describing protein sequences, polynucleotides, three-dimensional structures or literature, are generally defined by unambiguous identifiers also referred to as 'accession numbers'. The most prominent accession numbers also used in this work are the GenInfo Identifier (gi) used by the NCBI as well as the 'pdb' and 'sp' identifiers referring to entries in the PDB and SwissProt, respectively.

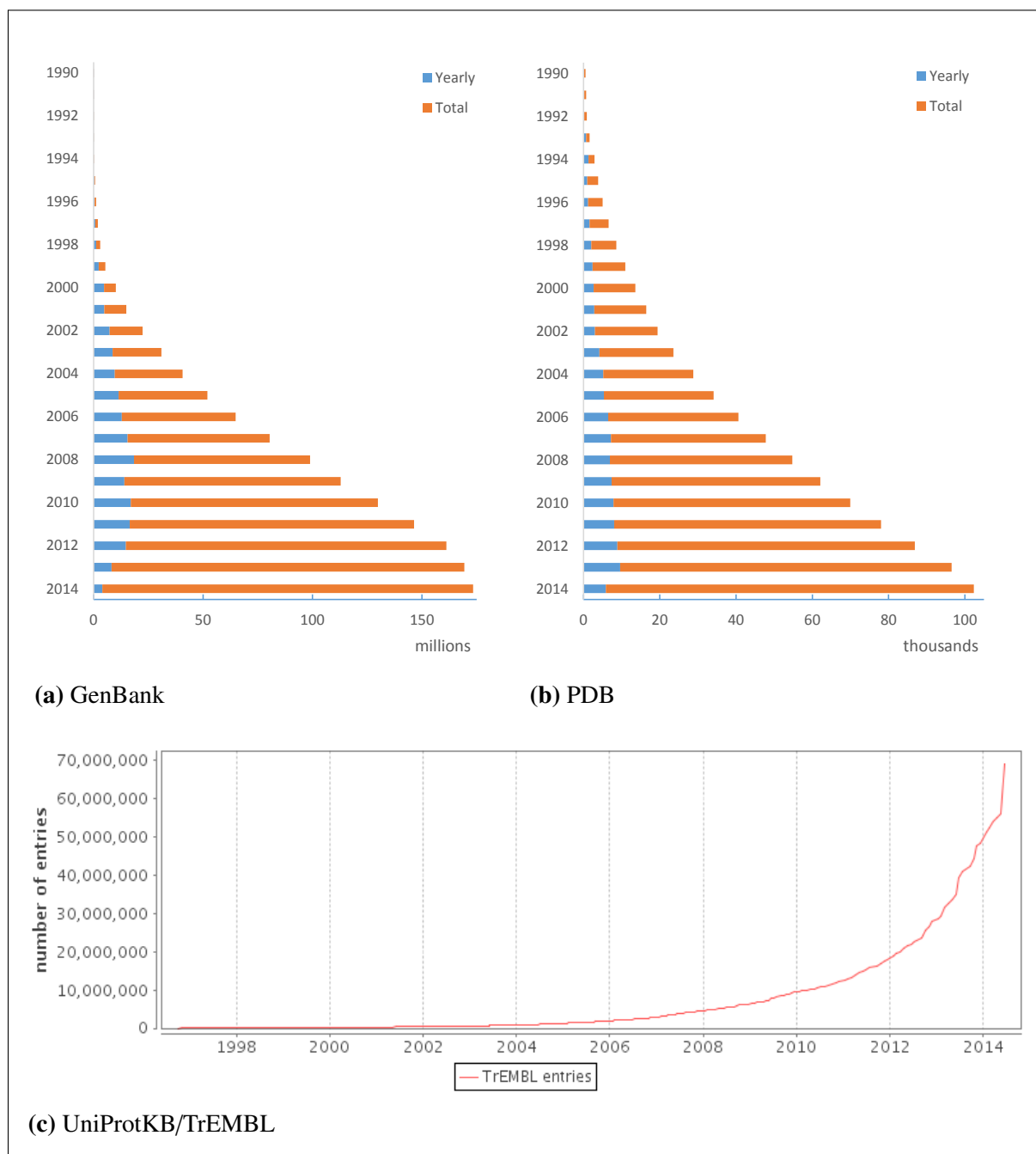


Figure 1.2: Statistics on the number of entries in the NCBI GenBank (Benson et al. 2011), the RCSB Protein Data Bank (PDB) (Berman et al. 2000) and UniProt/TrEMBL (The UniProt Consortium 2014) as of August 2014.

1.3 Family-specific protein databases (FSPDs)

Undoubtedly, there is a relationship between the sequence, the structure and the function of enzymes. The amino acid sequence determines the structure of an enzyme, although variations in the sequence do not implicitly provoke structural variations. Consequently, different enzymes with different amino acid sequences often share a highly similar three-dimensional structure. The function of an enzyme results from the correct placement of catalytically active molecules or groups and the appropriate orientation of substrates relative to them. Catalytic activity in enzymes can be exclusively conferred by the interplay of different amino acids, perfectly arranged in the enzyme's structure, as described for catalytic triads observed in a multitude of enzymes (Dodson and Wlodawer 1998). In contrast, cofactor-dependent enzymes, like the ThDP-dependent enzymes, possess the capability to bind additional, non-proteogenic molecules, which in turn provide catalytic activity. Consequently, the relationship between the sequence, the structure and the function of enzymes is rather specific for different enzyme families than a general rule for all enzymes. However, within protein families sharing a common reaction mechanism, these rules are most probably consistent. Thus, a systematic analysis of sequence-structure-function relationships has to be done family-specific.

Family-specific protein databases (FSPDs) are an appropriate basis for this kind of analysis, since

1. they organize comprehensive protein families, often consisting of tens of thousands of sequences in a manageable fashion,
2. they link sequence and structure information of proteins,
3. they provide information about functional relevance of distinct positions derived from experiments and
4. they give an overview over the amino acid composition of entire protein families.

All analyses done in this work were based on such FSPDs, which served as repositories for sequence and structure data, provided classification of subfamilies, and were in turn enriched with the derived results. In order to facilitate the generation and maintenance of FSPDs, systems capable to gather sequence and structure information for specific protein families were developed. In 2004, Markus Fischer² presented the *DWARF* (Data warehouse for analysing protein families)

²Markus Fischer, Dissertation 1998-2005, Institute of Technical Biochemistry, University of Stuttgart, Germany

system (Fischer 2004) and in 2007, Henk-Jan Jossten came up with the *3DM* system (Joosten 2007), both explicitly designed to generate and maintain such FSPDs, to serve as tools for navigation within families and to support rational engineering. While *3DM* was subsequently further developed into a commercial product, the *DWARF* system was kept in-house³.

Although originally developed by Markus Fischer to serve as the basis for analyses of α/β -hydrolases (Fischer 2004), the *DWARF* system evolved to the standard environment for database analyses at the Institute for Technical Biochemistry, University of Stuttgart, Germany. After being initially used to house and analyze the Lipase Engineering Database (LED) (Pleiss et al. 2000; Fischer and Pleiss 2003; Barth et al. 2004b), it was subsequently used and further optimized to house among others the CYtochrome P450 Engineering Database (CYPED) (Fischer, Knoll, et al. 2007; Sirim, Wagner, Lisitsa, et al. 2009), the Medium-Chain Dehydrogenase/Reductase Engineering Database (MDRED) (Knoll and Pleiss 2008), the PHA Depolymerase Engineering Database (DED) (Knoll, Hamm, et al. 2009), the LaCCase and multicopper oxidase Engineering Database (LCCED) (Sirim, Wagner, Wang, et al. 2011), the LACTamase Engineering Database (LACED) (Thai, Bös, and Pleiss 2009), the TriTerpene Cyclase Engineering Database (TTCED) (Racolta et al. 2012), the Metallo- β -Lactamase Engineering Database (MBLED) (Widmann, Pleiss, and Oelschlaeger 2012) and the Thiamine diphosphate-dependent Enzyme Engineering Database (TEED) (Widmann, Radloff, and Pleiss 2010). Thus, the fundamental relational data model and the systems for database generation, maintenance, curation and analysis of the *DWARF* system proved to be well suited for the investigation of sequence-structure-function relationships in different enzyme families. In order to comply with specific needs for different protein families, the system was continuously adopted. By optimizing the graphical user interface (GUI) of the *DWARF* system and implementation of object-oriented programming by Florian Wagner⁴, the performance of the GUI and the overall maintainability were increased. However, the algorithm of the basic parsing and pushing system, which was used

- to search for homologous sequences that belong to a defined protein family in online repositories,
- to 'parse' the available information on sequence, structure, annotations, source organisms and further data from those repositories,
- to 'push' this information into the FSPD,

³Institute of Technical Biochemistry, University of Stuttgart, Germany

⁴Florian Wagner, Institute of Technical Biochemistry, University of Stuttgart, Germany

- to classify subfamilies and
- to update existing databases,

was kept without changes significantly improving the parsing and pushing performance. As a consequence of the drastic increase in the number of publicly available amino acid sequences during the last ten years (2004-2014) (Figure 1.2 on page 4) and the accompanying increase in time-demand for generation and updates of FSPDs in the *DWARF* system, those algorithms were redesigned in this work to optimize the time-efficiency.

1.4 Homology modeling

Due to the often challenging process of generating protein crystals for X-ray crystallography and the experimental complexity of structure determination, precise structure information is available only for a small fraction of the known enzymes. However, investigation of the three-dimensional structure of an enzyme is highly beneficial for rational engineering, since it enables analysis of interactions between the enzyme and ligands, such as cofactors or substrates. Consequently, computational methods were developed in order to predict protein structures based on the respective amino acid sequence. One of those methods is homology modeling. It is based on the observation that homologous sequences likewise resemble each other in their structures (Browne et al. 1969) and that three-dimensional structures are more conserved within protein families than the amino acid sequences (Baker 2000). By modeling a sequence of interest (called 'target') guided by the structure of a homologous protein (called 'template'), atom coordinates can be predicted (Browne et al. 1969; Blundell et al. 1987). Thus, homology modeling starts with the search for templates with sequences homologous to the target sequence (Fiser and Sali 2003). By an alignment of the target and template sequences, regions with probable structural homology and regions more likely to structurally deviate are identified. For the former, the atom coordinates of the backbone atoms of the template structure are defined to be the coordinates of the corresponding atoms of the modeled target structure. Subsequently, loop modeling is applied to construct the backbone coordinates of the regions deviating between the target and the template sequences. The target structure is completed by modeling of the amino acid side chains. Usually, further quality estimation and model optimization follow, depending on the applied homology modeling tool. Different tools also differ in the methods used to generate and optimize

the initial target-template alignment, the strategy behind the loop modeling and the prediction of the side chain orientations.

1.5 Standard numbering schemes

A standard numbering scheme is a method to assign common numbers to corresponding positions of multiple homologous sequences. Standard numbering schemes were shown to be beneficial for research on protein families comprising sequences folding into similar three-dimensional structures. Establishment of such schemes for class A and B β -lactamases (Ambler et al. 1991; Galleni et al. 2001; Garau et al. 2004) and the complementary-determining regions of antibodies (Al-Lazikani, Lesk, and Chothia 1997; Kabat, Wu, and Perry 1991; Abhinandan and Martin 2008; Honegger and Plückthun 2001) improved communication on functionally relevant positions and facilitated identification of positions of interest in newly identified amino acid sequences. Even for protein families missing a designated standard numbering scheme, scientists refer to the position numbers of well-described sequences in the respective field to illustrate the location of discussed residues (O'Reilly, Watson, and Johnson 1999).

The need for standardized position numbers arises from deviating N-termini, which can be found even in highly similar sequences. Since numbering of amino acid positions in proteins generally starts at the respective N-terminus, homologous sequences exclusively deviating in the length of their N-termini differ in the numbers assigned to structurally equivalent positions. Consequently, comparison of scientific findings between such sequences is no straightforward task. In addition to the benefit in communication, standard numbering schemes have the potential to facilitate bioinformatical analysis. In order to analyze proteins for functional or structural relevant positions, commonly sequence alignments are generated and investigated for the amino acid composition at different positions. Amino acids arranged in the same column of a sequence alignment are interpreted as being located at the same position in the three-dimensional structures. In such analyses, multisequence alignments usually outperform the usage of pairwise alignments or single sequences, since they contain more information. Owing to the observation that only certain positions in homologous sequences are strictly conserved and that most positions underlie some mutational variation without negatively effecting the overall structure and function of the proteins, analysis of the variability of single positions allows to draw conclusions about their functional and structural relevance. Using a standard numbering scheme, all positions in

the sequences implicitly contain information about their structural correspondence due to the common standard position numbers. Thus, a systematic analysis of homologous sequences can be done without intermediate alignment steps after having applied a standard numbering scheme to all sequences of the respective protein family.

1.6 Thiamine diphosphate-dependent enzymes

ThDP-dependent enzymes form a vast and diverse family of proteins represented in all kingdoms of life. Comparable to their sequence diversity, ThDP-dependent enzymes are diverse in their catalytic potential. They catalyze a broad range of C-C bond cleavage and formation reactions. As it is common for enzymes, ThDP-dependent enzymes are enantioselective in the formation of various products, making them interesting tools in the synthesis of chiral building blocks. For instance, (*R*)-phenylacetylcarbinol ((*R*)-PAC) (Figure 1.3), which was first enzymatically produced by fermentation of yeast in the presence of benzaldehyde (Neuberg and Hirsch 1921), is used as a precursor of L-ephedrine (Hildebrandt and Klavehn 1930).

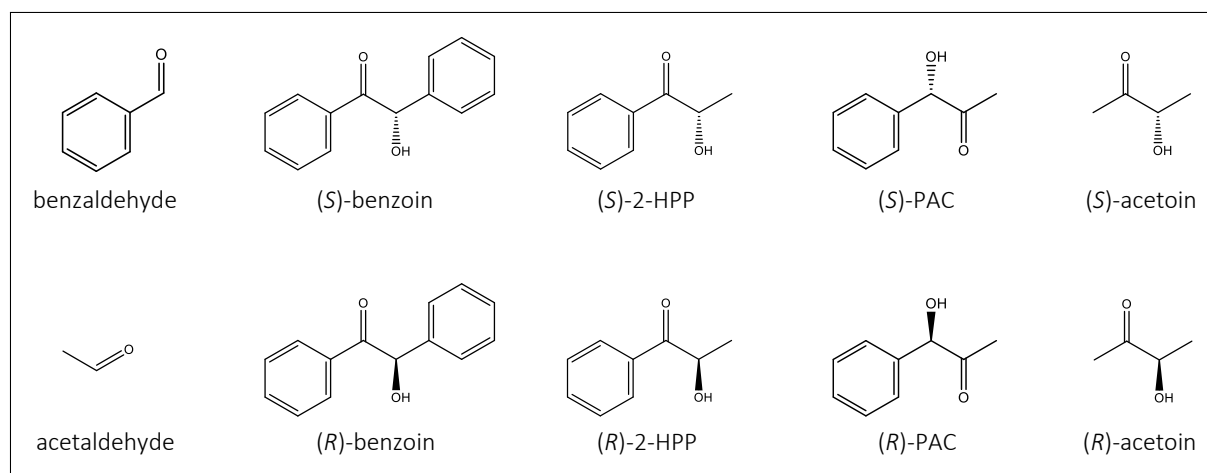


Figure 1.3: Benzaldehyde, acetaldehyde and the chiral products of homo- and crosscoupling. Homocoupling of benzaldehyde and acetaldehyde yields benzoin and acetoin, respectively. Crosscoupling yields 2-hydroxypropiophenone (2-HPP) and phenylacetylcarbinol (PAC).

By carbonylation of acetaldehyde and benzaldehyde, four different chimeric products are theoretically possible: condensation of two molecules acetaldehyde results in the formation of acetoin, condensation of benzaldehyde leads to formation of benzoin, and the mixed carbonylation yields 2-hydroxypropiophenone (2-HPP) or PAC, depending on the respective enzyme (Figure 1.4 on the following page).

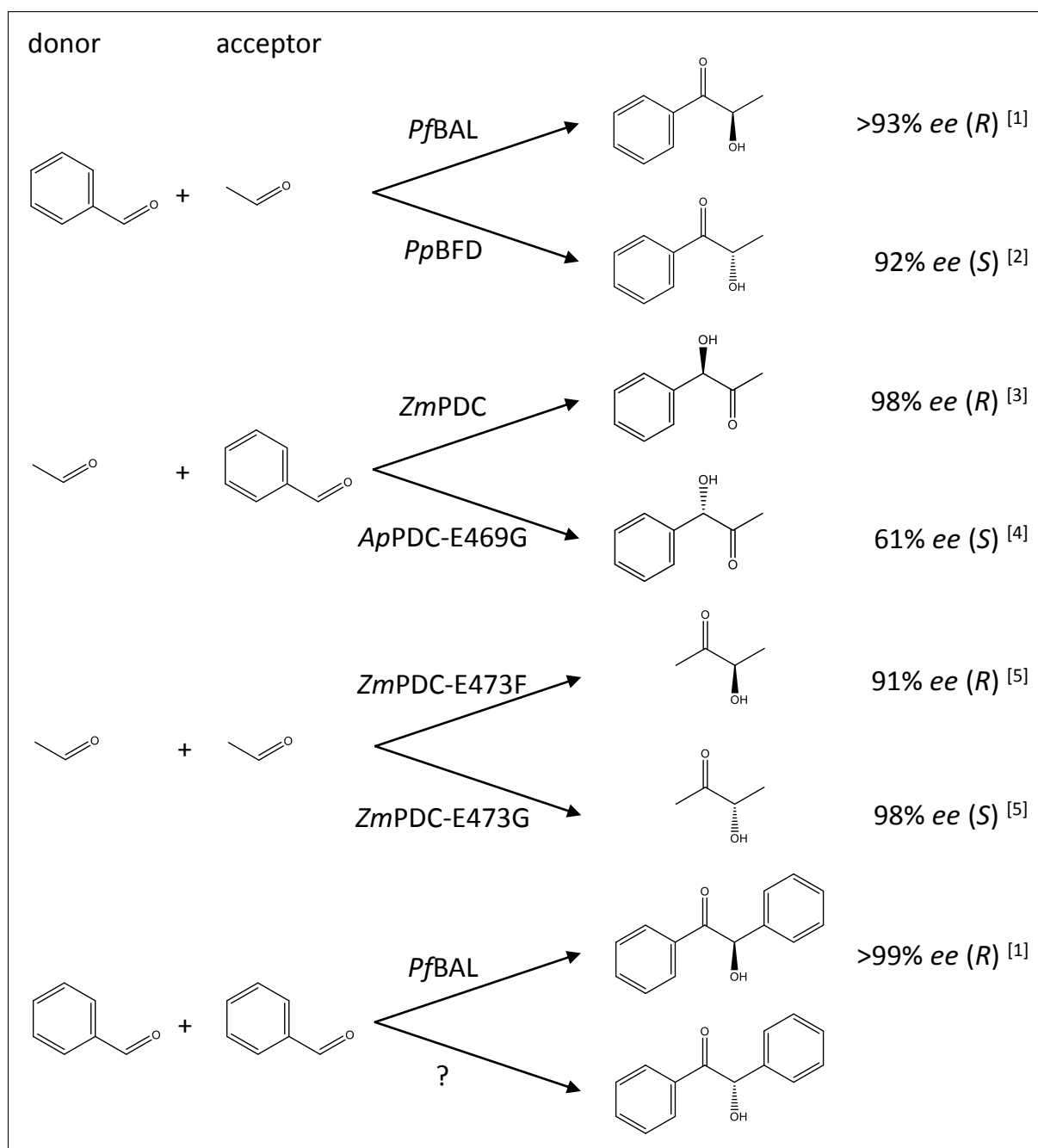


Figure 1.4: Via different ThDP-dependent enzymes, both enantiomers of 2-hydroxypropiophenone (2-HPP), phenylacetylcarbinol (PAC) and acetoin are accessible starting from benzaldehyde and acetaldehyde. *PfBAL* catalyzes the synthesis of (*R*)-2-HPP and additionally provided access to enantiopure (*R*)-benzoin ([1], Demir, Sesenoglu, Eren, et al. 2002). (*S*)-2-HPP is formed in the crosscoupling reaction by *PpBFD* ([2], Iding, Dünnwald, et al. 2000). *ZmPDC* catalyzes the crosscoupling of both substrates to (*R*)-PAC ([3], Siegert et al. 2005), while the variant *ApPDC-E469G* predominantly catalyzes the formation of (*S*)-PAC ([4], Rother et al. 2011)). (*R*)- and (*S*)-acetoin are accessible via the variants *ZmPDC-E473F* and *ZmPDC-E473G*, respectively ([5], Wechsler 2014). Only (*S*)-benzoin was not accessible by carbonylation of benzaldehyde using the available enzyme platform so far.

Using different ThDP-dependent enzymes and engineered variants thereof, almost the entire scope of possible products in reactions starting from acetaldehyde and benzaldehyde was accessible (Figure 1.4). The benzaldehyde lyase from *Pseudomonas fluorescens* (*PfBAL*) was shown

to almost exclusively catalyze the formation of (*R*)-2-HPP in the mixed carbonylation of both substrate aldehydes and (*R*)-benzoin in the homocoupling of benzaldehyde (Demir, Sesenoglu, Eren, et al. 2002). (*R*)-PAC and (*R*)-acetoin are accessible via heterocoupling by the pyruvate decarboxylase from *Zymomonas mobilis* (*ZmPDC*) (Siegert et al. 2005) and homocoupling of acetaldehyde by the *ZmPDC*-E473F variant (Wechsler 2014), respectively. Benzoylformate decarboxylase from *Pseudomonas putida* (*PpBFD*) predominantly catalyzes the formation of (*S*)-2-HPP (Iding, Dünwald, et al. 2000), (*S*)-PAC and (*S*)-acetoin became accessible via variants of the pyruvate decarboxylase from *Acetobacter pasteurianus* (*ApPDC*-E469G, Rother et al. 2011) and *ZmPDC* (*ZmPDC*-E473G, Wechsler 2014), respectively. Only the direct asymmetric synthesis of (*S*)-benzoin starting from benzaldehyde was not possible using the available platform of ThDP-dependent enzymes, so far (Figure 1.4 on the preceding page). By rational engineering, an *ApPDC* variant was designed in this work, able to close this gap.

1.6.1 Thiamine diphosphate

ThDP is the activated form of thiamine (Vitamin B₁) and is produced by an ATP:thiamine diphosphotransferase (also called thiaminokinase) (Kaziro et al. 1961). ThDP is a cofactor in various enzymatic reactions including key steps in energy metabolism, preparation of reducing equivalents and reaction intermediates. The E1 component of the pyruvate dehydrogenase complex, an enzyme utilizing ThDP as a cofactor, is involved in the conversion of pyruvate coming from glycolysis into acetyl-CoA to supply the citric acid cycle (Berg, Tymoczko, and Stryer 2014). Transketolase, likewise depending on the catalytic potential of ThDP, plays a key role in the pentose phosphate pathway and the Calvin cycle (Kochetov and Solovjeva 2014). Thiamine, which consequently is essential for the metabolism of all organisms, is synthesized by bacteria, fungi and plants and has to be ingested by animals and humans. Its essential role as a nutrient is indicated by deficiency symptoms like Wernicke-Korsakoff's syndrome and Beriberi resulting from insufficient intake with the diet (Zbinden 1962). Latest research further suggests a role of thiamine in Alzheimer's disease (Lu'o'ng and Nguyen 2011; Gibson et al. 2013). ThDP consists of a pyrimidine and a thiazole moiety, the latter comprising the catalytically relevant C2 atom (Breslow 1957), and a pyrophosphate group (Figure 1.5 a on the following page).

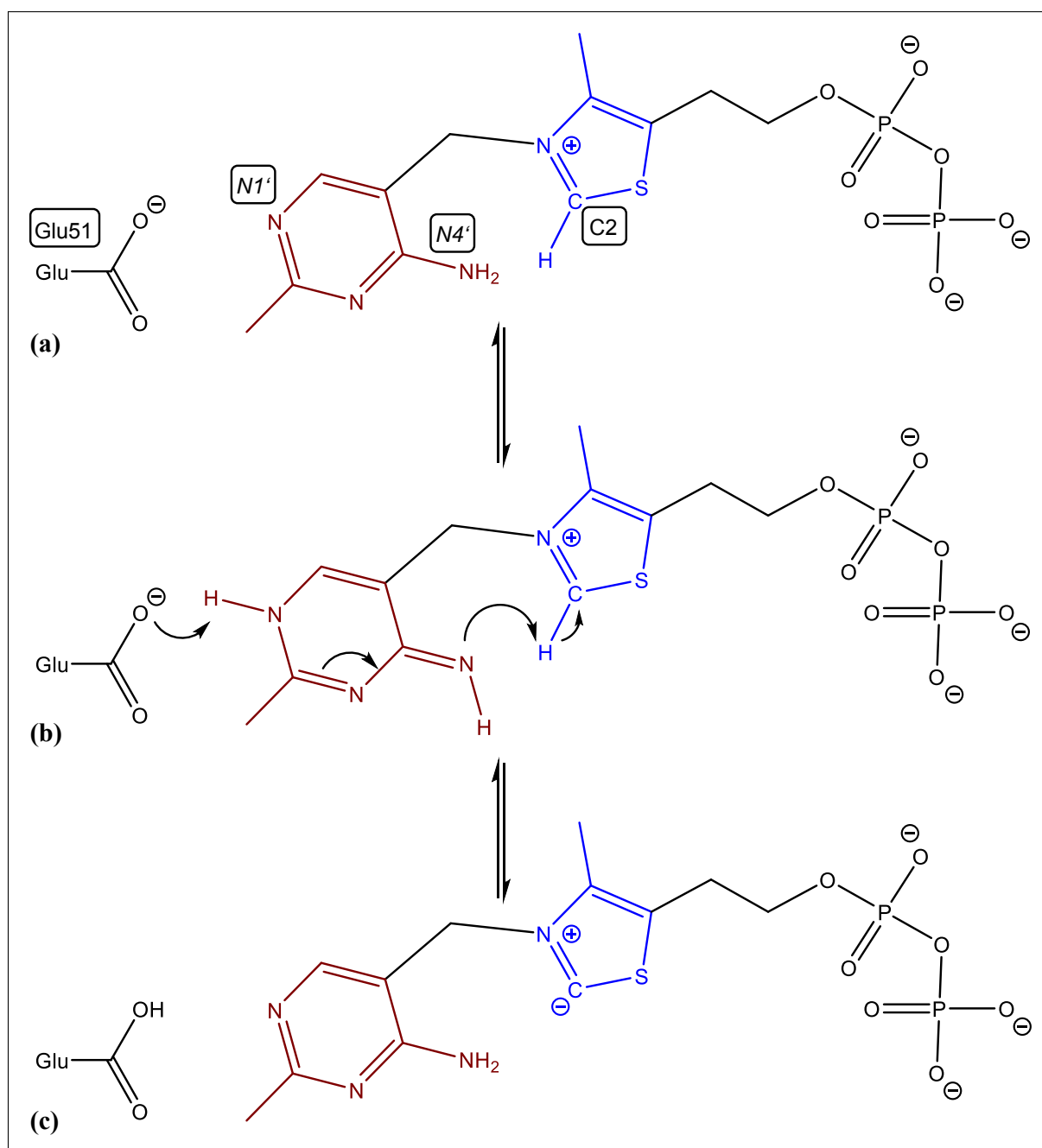


Figure 1.5: Thiamine diphosphate consists of an aminopyrimidine (red), a thiazole (blue) and a pyrophosphate moiety (a). The aminopyrimidine can tautomerize from the amino (a) to the imino form (b) (Meyer, Neumann, Ficner, et al. 2013). In most ThDP-dependent enzymes, the cofactor is subsequently activated by deprotonation at the N1' atom of the pyrimidine ring by a conserved glutamate residue (at standard position 51 referring to the standard numbering scheme for ThDP-dependent enzymes presented in this work, see Section 2.2.1 on page 32). As a consequence of the 'V-conformation', in which the cofactor is constrained in all ThDP-dependent enzymes, the C2 atom in the thiazole ring can subsequently be deprotonated by the 4-imino group of the aminopyrimidine (b). The resulting ylide form represents the activated cofactor (c). The methylene bridge linking the aromatic rings of the thiazole and the pyrimidine was stretched in order to allow for the 2D-representation of the neighboring C2 and N4' atoms.

1.6.2 Reaction mechanism

Catalysis in ThDP-dependent enzymes, although they are diverse in their sequences, the substrates and specific products, follows a common mechanism. In order to allow for its catalytic function, the ThDP cofactor has to be activated by the protein. Therefore, ThDP-dependent enzymes constrain the cofactor into a 'V'-shaped conformation and with exception of glyoxylate carboligases (Kaplun et al. 2008), a conserved glutamate plays a key role in the activation. The so called 'V-conformation', which juxtaposes the 4'-imino group of the aminopyrimidine and the C2 atom of the thiazolium moiety (Jordan 2003; Frank, Leeper, and Luisi 2007; Kaplun et al. 2008), was observed for all ThDP-dependent enzymes with experimentally determined structure (Shin et al. 1977; Pletcher et al. 1977; Dobritsch et al. 1998; Guo et al. 1998; Pang et al. 2004; Knoll, Müller, et al. 2006; Xiang et al. 2007; Shaanan and Chipman 2009; Werther et al. 2010; Andrews and McLeish 2012; Meyer, Neumann, Koers, et al. 2012). Activation is initiated by abstraction of a proton from the N1' atom by the conserved glutamate residue, which finally results in deprotonation of the C2 atom by the 4'-imino group of the aminopyrimidine (Figure 1.5 on the preceding page). The activated cofactor in the ylide⁵ form is subsequently attracted by electrophilic groups, such as carbonyls of aldehydes or α -keto acids. Binding of the substrate ('donor') to the C2 atom of the ThDP results in bond breaking, which is a deprotonation or decarboxylation in the case of aldehydes or α -keto acids, respectively (Figure 1.6 on the next page). The resulting metastable carbanion-enamine intermediate (referred to as 'Breslow intermediate') can be attacked by a second substrate ('acceptor') leading to formation of an aldehyde, in the case of a proton as the acceptor substrate, or an α -hydroxy ketone, in the case of an aldehyde as acceptor. By protonation via a proton relay system, the ThDP ylide is regenerated (Frank, Leeper, and Luisi 2007).

1.6.3 Regulation of enantio- and chemoselectivity in ThDP-dependent enzymes

Starting from the acetaldehyde and benzaldehyde, eight different products can be formed by carbonylation (Figure 1.4 on page 10). Since ThDP-dependent enzymes selectively catalyze the formation of either PAC or 2-HPP (and derivatives thereof) or at least show preferences in the

⁵Recent results obtained from high-resolution X-ray crystallography and near-UV CD spectroscopy further indicate existence of the carbene form of deprotonated ThDP in *LpPOX* (Meyer, Neumann, Ficner, et al. 2013).

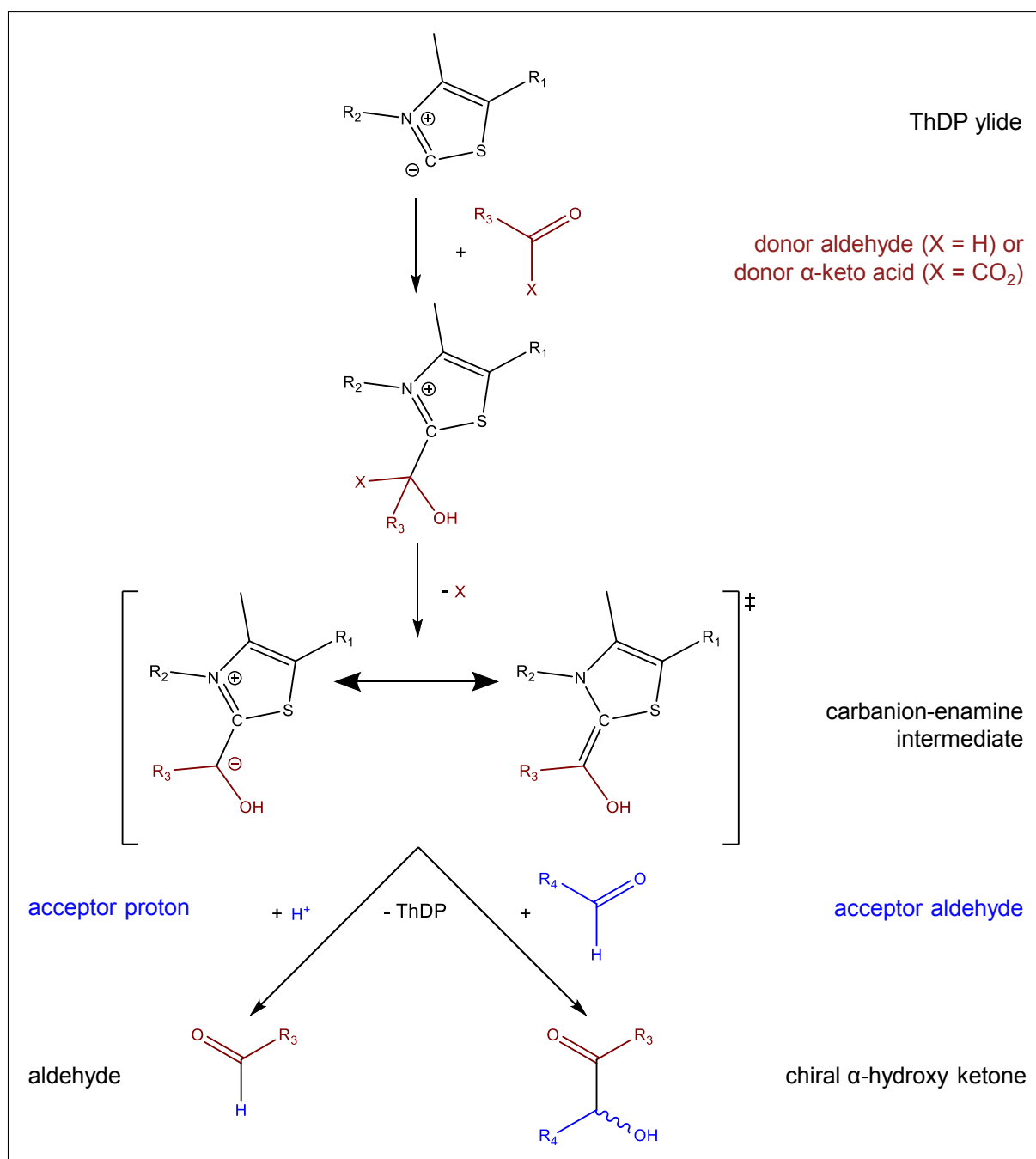


Figure 1.6: In the reaction mechanism of ThDP-dependent enzymes (according to Schellenberger 1998), the ThDP ylide is first attacked by the donor aldehyde or α -ketoacid. The donor substrate is deprotonated or decarboxylated, respectively, resulting in the formation of a carbanion-enamine intermediate. Electrophilic attack by a proton yields the product aldehyde, whereas attack by an aldehyde leads to C-C bond formation and release of the corresponding α -hydroxy ketone product.

mixed carbonylation, the enzymes must possess mechanisms steering chemoselectivity. Further, the enzymes control the selective formation of preferred enantiomers of the chiral products. Via detailed investigation of the structural differences of *PfBAL* and *PpBFD*, Knoll and co-workers identified a small pocket in the active site of *PpBFD* (Knoll, Müller, et al. 2006) (Figure 1.7 a/b).

The so-called *S*-pocket, which exists in *Pp*BFD but is missing in *Pf*BAL, was assumed to allow acetaldehyde to bind in an antiparallel arrangement relative to the ThDP-bound benzaldehyde prior to carboligation. As a result, *Pp*BFD predominantly catalyzes the formation of (*S*)-2-HPP, while the same reaction catalyzed by *Pf*BAL yields (*R*)-2-HPP.

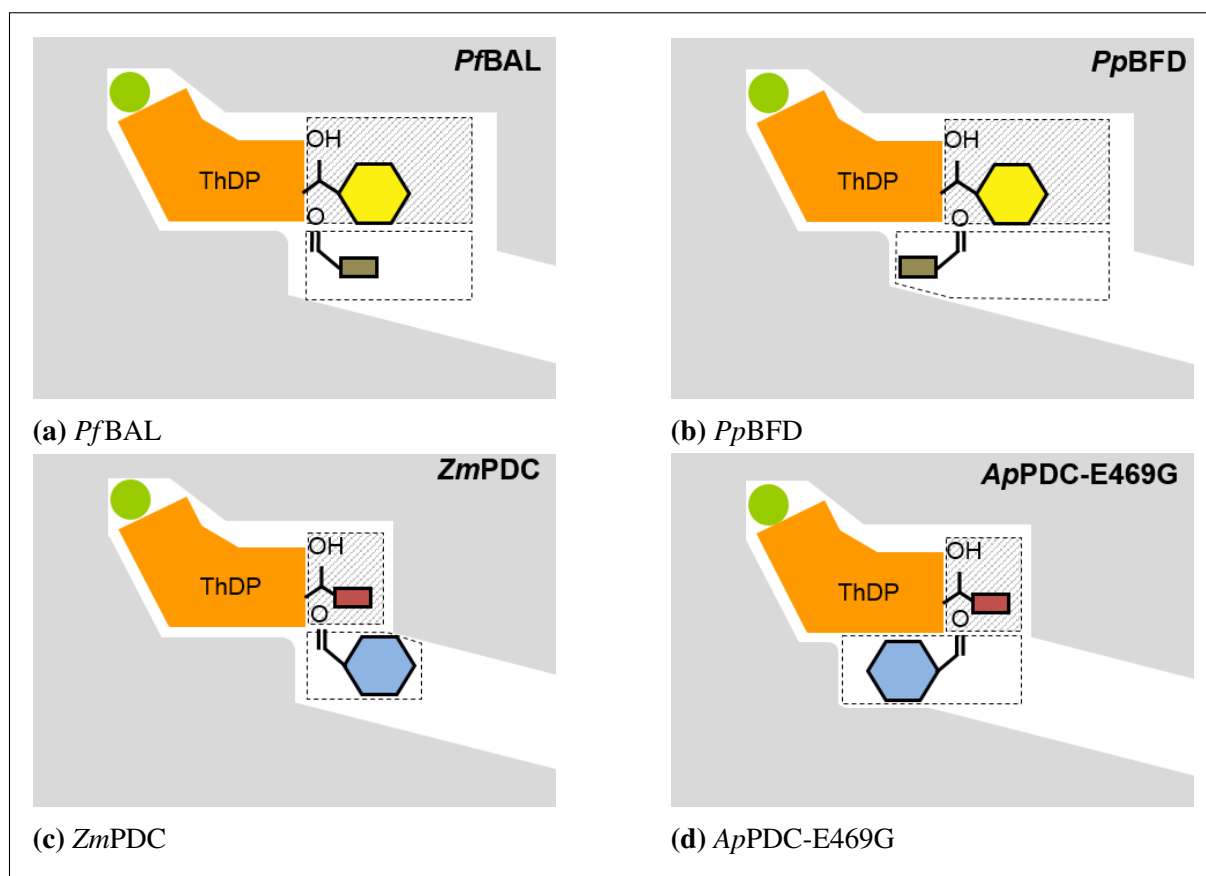


Figure 1.7: The schematic representation of the donor and acceptor binding sites of *Pf*BAL, *Pp*BFD, *Zm*PDC and *Ap*PDC-E469G (derived from Westphal 2013) reveals the different binding modes of acetaldehyde and benzaldehyde resulting in different products in the mixed carboligation. The different chiral products are shown in Figure 1.4 on page 10. Owing to the parallel arrangement of donor benzaldehyde and acceptor acetaldehyde in *Pf*BAL, the heterocoupling yields (*R*)-2-HPP. Due to the widened acceptor binding site forming a *S*-pocket, which stabilizes acetaldehyde in an antiparallel arrangement relative to the donor substrates, *Pp*BFD catalyzes the formation of (*S*)-2-HPP. (*R*)-PAC is formed by *Zm*PDC, since the small donor-binding site preferably stabilizes acetaldehyde and the missing *S*-pocket constrains the acceptor benzaldehyde into a parallel arrangement. By enlargement of the acceptor-binding site (*S*-pocket) in *Ap*PDC-E469G (Rother et al. 2011), the antiparallel arrangement is allowed, resulting in the formation of (*S*)-PAC. Red rectangle: donor acetaldehyde, brown rectangle: acceptor acetaldehyde, yellow hexagon: donor benzaldehyde, blue hexagon: acceptor benzaldehyde, green dot: Mg^{2+} ion, ruled box: donor-binding site, white box: acceptor-binding site.

Thus, the *S*-pocket concept explains the enantioselectivity of ThDP-dependent decarboxylases. However, this enzyme family further controls the chemoselectivity. Pohl and co-workers explained the preference in the formation of 2-HPP or PAC by different shapes of the binding pockets for the donor and acceptor substrates in mixed carboligations (Pohl, Gocke, and Müller 2010). Binding of benzaldehyde to the ThDP cofactor as the donor substrate and subsequent

carbonylation with acetaldehyde in the acceptor position leads to formation of 2-HPP (Figure 1.7 a/b), whereas the reverse combination of acetaldehyde as the donor and benzaldehyde as the acceptor substrate yields PAC (Figure 1.7 c/d). Consequently, the size and stabilizing effects of the donor pocket are relevant for the chemoselectivity, whereas the size and substrate stabilization of the acceptor-binding site (*S*-pocket) determine the enantioselectivity.

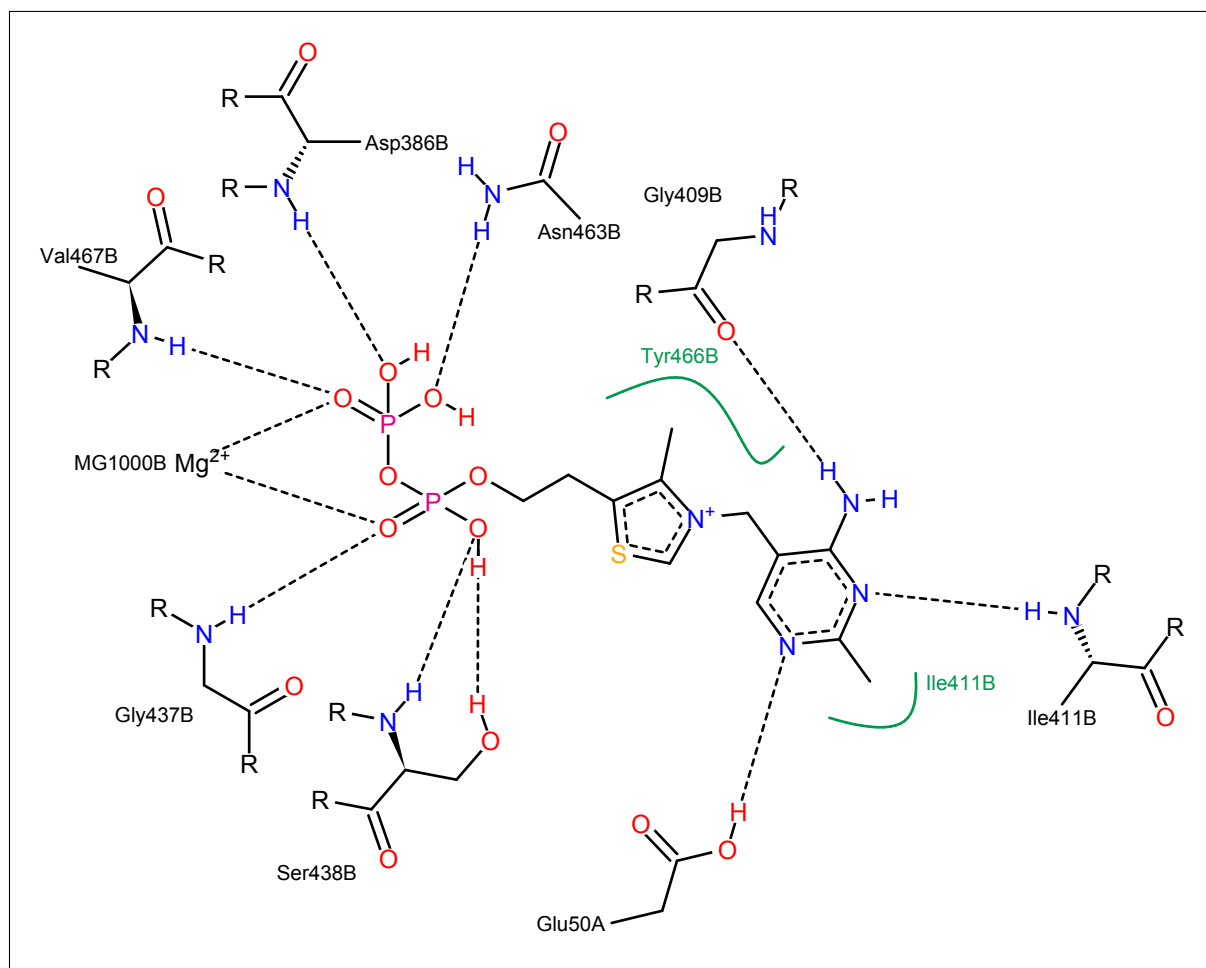


Figure 1.8: Schematic representation of the binding of ThDP in one of the active sites of *ApPDC*. Glycine 437, serine 438 and asparagine 463 form part of the **GDGX_{24,27}N**-motif (bold residues). The glutamic acid at position 50 of *ApPDC* (standard position 51 according to the standard numbering scheme for ThDP-dependent decarboxylases (Vogel, Widmann, et al. 2012)) is the key residue involved in activation of the cofactor (Figure 1.5 on page 12). The diagram was constructed using the 3D-coordinates of the *ApPDC* and the ThDP cofactor as deposited in the Protein DataBank (pdb|2VBI, Rother et al. 2011) and *PoseView* (Stierand and Rarey 2007) for calculation of the interactions and drawing. The naming of the residues corresponds to the position numbers as defined in pdb|2VBI. The letters A and B, which complement the residue names, indicate the respective monomers of the homodimer forming the protein structure. Black dashed lines: hydrogen bonds, salt bridges and metal interactions; green lines: hydrophobic interactions.

1.6.4 Sequence and structure of ThDP-dependent enzymes

Although diverse in sequence and structure, all ThDP-dependent enzymes share common features like the binding of the cofactor buried within the protein. Binding of ThDP is facilitated by two domains called pyrimidine (PYR) and pyrophosphate (PP) binding domains, which contact the cofactor at the pyrimidine and pyrophosphate moieties, respectively (Duggleby 2006). In all ThDP-dependent enzymes, the PP domain coordinates a divalent cation (Mg^{2+} , Mn^{2+} or Ca^{2+}), which in turn binds the cofactor by electrostatic interactions with the negatively charged pyrophosphate group (Dobritzsch et al. 1998). In addition, the PP domains encompass highly conserved residues described as the GDGX_{24,27}N-motif relevant for ThDP binding (Hawkins, Borges, and Perham 1989). Another well-described, highly conserved position in most ThDP-dependent enzymes is the glutamate residue responsible for the cofactor activation. Located in the PYR domain, it points towards the pyrimidine moiety of ThDP (Figure 1.8 on the preceding page).

Besides the PYR and PP domains, responsible for binding and arranging the ThDP cofactor in the 'V-conformation', ThDP-dependent enzymes often comprise further domains. Characteristic for members of the decarboxylase superfamily of ThDP-dependent enzymes is the presence of a so-called TH3 domain (Figure 1.9 on the following page), bearing its name due to structural similarity with the dIII component of the transhydrogenase from *Rhodospirillum rubrum* (Duggleby 2006). Similar to the transhydrogenase dIII component, the TH3 domain of ThDP-dependent decarboxylases conveys binding of nucleotides. However, this function was maintained only in pyruvate oxidases (POX) (Tittmann, Wille, et al. 2005), oxalyl-CoA decarboxylases (OCDC) (Berthold, Moussatche, et al. 2005), glyoxylate carboligases (GXC) (Chung, Tan, and Suzuku 1971), cyclohexane-1,2-dione hydrolase (CDH) (Steinbach et al. 2012), and acetohydroxy acid synthases (AHAS) (Lee, Lee, et al. 2013; Duggleby and Pang 2000; McCourt et al. 2006), while it lost its original capability to bind nucleotides like flavin adenine dinucleotide (FAD) or adenosine diphosphate (ADP) in other subfamilies. For the family of transketolases a C-terminal domain (TKC) was described to complement the PYR and PP domain (Costelloe, Ward, and Dalby 2008), as well as there are additional domains in ThDP-dependent oxidoreductases (Chabrière et al. 1999) and 2-oxoglutarate decarboxylases (Wagner et al. 2011). Frank and co-workers described all ThDP-dependent enzymes to function as dimers with two active sites or tetramers with four active sites (Frank, Leeper, and Luisi 2007).

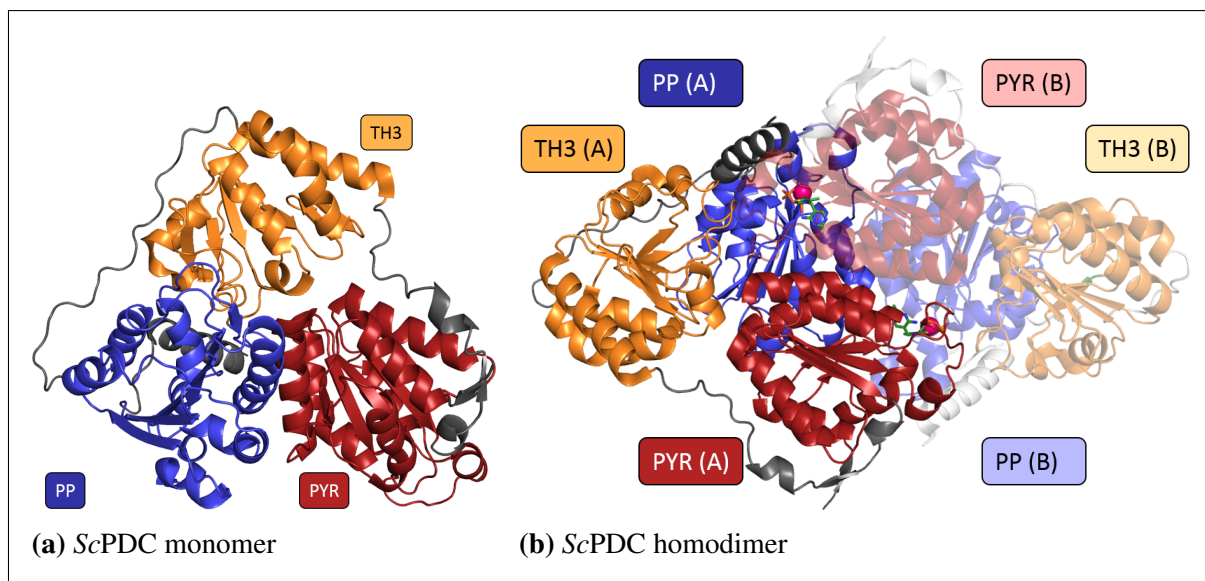


Figure 1.9: Each monomer of the pyruvate decarboxylase from *S. cerevisiae* (*ScPDC*) is composed of an N-terminal PYR domain (red), a C-terminal PP domain (blue) and a linking TH3 domain (orange) (a). Assembly of two such monomers (referred to as monomers A and B) results in formation of the active dimer (b). In the dimer, two active sites containing each one molecule of ThDP (green sticks) and a Mg^{2+} ion (pink sphere) are formed at the interface of PYR and PP domains of different monomers.

1.6.5 Family classification

Based on sequence and structure similarity, ThDP-dependent enzymes were classified into different subfamilies. In 2006 Duggleby discussed seven different subfamilies of ThDP-dependent enzymes with characteristic compositions of domains and deviating structures: sulfopyruvate decarboxylases (SPDC), phosphonopyruvate decarboxylases (PPDC), decarboxylases (DC), transketolases (TK), two families of α -ketoacid dehydrogenases (K1 and K2) and the pyruvate:ferredoxin oxidoreductases (Duggleby 2006). In 2007 Costelloe and co-workers discussed six families with deviating domain arrangement. Based on this criterion, they separated the family of ThDP-dependent enzymes into pyruvate decarboxylase-like, transketolase-like, pyruvate:ferredoxin oxidoreductases, 2-oxoisovalerate dehydrogenase-like enzymes, SPDCs and PPDCs (Costelloe, Ward, and Dalby 2008). In 2010 Widmann and co-workers suggested to classify ThDP-dependent enzymes into eight superfamilies with deviating domain arrangement and functional annotation (Widmann, Radloff, and Pleiss 2010), combining the findings of Costelloe, Ward, and Dalby 2008 and Duggleby 2006. This classification into decarboxylases, transketolases, oxidoreductases, two families of α -ketoacid dehydrogenases (K1 and K2), SPDC, PPDC and α -ketoglutarate dehydrogenases served as the basis for the creation of the ThDP-dependent Enzyme Engineering Database (TEED) (Widmann, Radloff, and Pleiss 2010).

1.6.6 ThDP-dependent Enzyme Engineering Database (TEED)

The TEED was established in 2010 within the *DWARF* system for the generation of FSPDs (Widmann, Radloff, and Pleiss 2010). Using seed sequences representing the eight superfamilies, 12048 sequences of 9443 different proteins were collected and incorporated in the TEED. 63 homologous families were established within the superfamilies to further classify sequences based on sequence similarity. Homologous families were named according to the descriptions of the corresponding entries in the online repositories that served as the sources for the database generation. Multisequence alignments, phylogenetic trees and Hidden Markov model (HMM) profiles were provided for visual inspection and download via an online accessible version⁶. Further, the PYR and PP domains were identified and annotated within the 63 homologous families. In this work a new, more comprehensive version of the TEED was generated based on the sequence information and family classification provided by Widmann and co-workers.

⁶www.teed.uni-stuttgart.de - Site was replaced by an updated version established as part of this work.

1.7 Scope and objectives





ThDP-dependent enzymes are able to catalyze the chemically challenging formation of C-C bonds. In combination with their enantioselectivity in the formation of chiral products, they are well-suited for an application in biocatalysis. Moreover, rational engineering enlarged the toolbox of ThDP-dependent enzymes with excellent enantioselectivities for non-physiological substrates (Pohl, Gocke, and Müller 2010). In order to support rational engineering of ThDP-dependent enzymes and to facilitate design of variants with desired characteristics, this thesis was intended to investigate the relationships between the sequences, structures and documented functions of this enzyme family. The title "Systematic analysis of the sequence-structure-function relationships of ThDP-dependent enzymes" adumbrates the two major aspects of this thesis: The *systematic analysis* and the *sequence-structure-function relationships of ThDP-dependent enzymes*. Methods for the systematic analysis of protein families were developed and subsequently applied in order to investigate the mentioned relationships in the family of interest. In addition, rational engineering was applied to design enzyme variants capable to catalyze the asymmetric carbonylation of benzaldehyde to (*S*)-benzoin. For this purpose, the following tasks were performed:

- The TEED was intended to serve as the basis for the analysis of the sequence-structure-function relationships of ThDP-dependent enzymes. Therefore, it should be updated to optimally represent the entire protein family at the time of the analysis. However, due to the steep increase in the number of known and publicly available protein sequences (Figure 1.2 on page 4), the *DWARF* system used for the initial generation of the TEED was no more able to keep this database up-to-date. Thus, a performance-optimized toolbox for the generation and maintenance of FSPDs was developed in order to keep up with the increasing number of known sequences.
- Using the developed toolbox, the TEED was updated in order to re-enable a systematic analysis of this diverse protein family.
- To facilitate a systematic analysis of the relationships between sequences, structures and functions of ThDP-dependent enzymes, information on all three components has to be made electronically available. Although sequence and structure information was previously successfully incorporated in the TEED, information on the biochemical behavior of the respective biocatalysts was missing. Thus, a new system for the generation, maintenance

and analysis of FSPDs capable to combine all three components was developed.

- A method for the identification of corresponding positions in different ThDP-dependent enzymes was required in order to facilitate the analysis of functional roles of distinct positions. Existing tools for the generation of multisequence and structure alignments provided methods for this purpose but they rely on the availability of multiple homologous sequences or structures at the time of the analysis. To allow the identification of positions of interest on any sequence belonging to the family of ThDP-dependent enzymes without the need for additional steps of homology search and alignments, a method for the generation of standard numbering schemes was developed.
- Sequences and structures of ThDP-dependent enzymes should be systematically analyzed to identify subfamily-specific differences. Due to the deviating domain composition in different superfamilies, this analysis had to be focused on the conserved PYR and PP domains. An investigation of the available structures combined with an analysis of the amino acid sequences of ThDP-dependent enzymes revealed simple principles in their modular construction. Those findings furthermore enabled the design of a synthetic construct with altered domain architecture and putatively altered structural composition.
- Systematic analysis of the amino acid distribution within subfamilies, combined with detailed investigation of structural differences between functionally different enzymes, enabled rational engineering of enzymes with desired chemo- and enantioselectivities.

2 Results

Due to the broad scope of this thesis' objectives including method development, application of existing and custom-built methods as well as interpretation of biological observations, the following chapter describes a mixture of newly developed methods and tools as well as the biological findings acquired by their application to the family of Thiamine diphosphate (ThDP)-dependent enzymes. Parts of this thesis have already been published¹ and the four publications most relevant for the presented results are appended to this work (Chapter 4 on pages 76ff.). References to figures, tables or further information in those publications are labeled with , ,  and , respectively.


2.1 Generation and maintenance of family-specific protein databases

2.1.1 The Thiamine diphosphate-dependent Enzyme Engineering Database

In 2010, Widmann and co-workers presented the ThDP-dependent enzyme engineering database, also referred to as TEED (Widmann, Radloff, and Pleiss 2010). Integrated into the *DWARF* system (Fischer 2004; Fischer, Thai, et al. 2006), it was intended to serve for the analysis of ThDP-dependent enzymes. Due to the permanent increase in the numbers of known sequences and structures, family-specific protein databases (FSPD) like the TEED have to be updated on a regular basis in order to effectively represent the known sequence and structure space. Consequently, the TEED was updated as a part of the work presented here. Due to limitations in the performance and the provided features in the *DWARF* system, novel tools for the generation and

¹For a complete list of all publications deduced from this thesis see page 229

maintenance of FSPDs were developed (*DBParse*, Section 2.1.2 and *DBUpdate*, Section 2.1.3). Furthermore, the underlying data model of the *DWARF* system was renewed and extended in order to enable depiction of functional information besides sequence and structure information of proteins. In consequence, the *DWARF* system was substituted by a new system called *BioCatNet* (Section 2.1.4 on page 28).

The most recent version of the TEED was generated and updated using *DBParse* and *DBUpdate* and was implemented in *BioCatNet* (www.teed.biocatnet.de). The family classification was based on the previous version of the TEED but extended by an additional superfamily, and 105 additional homologous families. The 1-Deoxy-D-xylulose-5-phosphate synthases (DXPS), previously classified as a homologous family belonging to the transketolase superfamily, were separated due to characteristic differences in the structural architecture (for details see Section 2.3.1 on page 40 and  Section 4.2 on pages 92ff.). With the establishment of *BioCatNet*, two further hierarchical levels (superfamily groups (SFAM groups) and groups of homologous families (HFAM groups)) were introduced. By combining the DXPS and transketolase (TK) superfamilies in a 'TK-like' SFAM group, the sequence similarity between both families was represented. By the use of HFAM groups, densely populated superfamilies like the decarboxylases (DC) and TKs were further classified into groups of proteins with putatively similar function based on the protein names. In addition, individual homologous families were grouped by the different subunits in case of heteromultimeric enzymes (e.g. the α and β subunits of sulfopyruvate decarboxylases). The TEED was furthermore enriched by information specific for the different super- and homologous families. Findings from an analysis of the modular structure of ThDP-dependent enzymes (see Section 2.3.1 on pages 40ff.) were summarized and implemented as additional information besides data on the average sequence length, the distribution of sequences and structures in different families, and the taxonomic origins of the encompassed sequences.

2.1.2 DBParse

The *DWARF* system established by Markus Fischer (Fischer 2004; Fischer, Thai, et al. 2006) was used to create, maintain and analyze a variety of different FSPDs over the last 14 years. However, due to the drastic increase in the number of available sequences and structures of proteins from all families (Figure 1.2 on page 4), the generation of new FSPDs or updates of existing databases became an interminable task. In order to catch up with the development in the number

of available information on proteins, a new toolbox called *DBParse* was developed. Partially reusing the existing programs for parsing XML files and inserting data into the database from the *DWARF* system, the newly developed *DBParse* makes use of parallelization to accelerate database generation. As inherited from the *DWARF* tools, *DBParse* uses representative 'seed' sequences to define a protein family. Those seed sequences are used as queries in *BLAST* (Camacho et al. 2009) searches against the non-redundant protein database provided by the NCBI (Benson et al. 2011) with the objective to identify all homologous sequences belonging to the respective protein family. *DBParse* splits the database generation process in six substeps, which are subsequently executed:

1. Searching for homologous sequences,
2. download of the XML files for all identified homologs,
3. sorting of sequences into non-redundant sequence and protein entries,
4. sorting into homologous families,
5. parsing of additional information and pushing of the presorted data into the database,
6. as well as searching for experimentally determined protein structures and integration into the database.

As compared to the *DWARF* system, each step is applied to the full set of sequences before execution of the subsequent steps. In the previous tools of the *DWARF* system, each seed sequence was used as a query for a *BLAST* homology search against a local copy of the non-redundant protein database provided by the NCBI. All identified homologs were subsequently processed by sequentially downloading the respective XML file from the NCBI protein database and execution of the parsing and pushing algorithm before the next seed sequence was used for the next iteration. Thus, temporary disconnections between the local machines and the NCBI servers led to eventually undocumented skips of sequences and structures.

In addition to the more failsafe strategy, the stepwise execution chosen for *DBParse* allows for optimal exploitation of the potential to parallelize the database generation process. *DBParse* starts multiple *BLAST* searches for the different seed sequences in parallel. Controlled by an implemented queuing system, optimal distribution of the different jobs to idle CPU cores is ensured. In addition, implementation of *blastp* (Camacho et al. 2009), an performance-optimized version of the protein *BLAST* application, accelerates the search for homologous sequences.

After the identification of all homologs, the corresponding XML files are fetched from the NCBI servers. Failed downloads are documented and automatically restarted in order to receive the correct data for each found sequence entry. Subsequently, the amino acid sequences are read from the XML files, sequences shorter than a minimum length (default: 150 amino acids) are discarded and the remaining sequences are sorted to sequence and protein entries using the *cluster_fast* algorithm provided within the *usearch* program (Edgar 2010). Sequences and fragments thereof are identified and subsequently regarded as combined sequence entries. In contrast to the previous implementation, identical amino acid sequences originating from different source organisms are treated as the same sequence entry. Furthermore, sequences with global identities higher than a defined threshold (default: 98%) are combined to protein entries, independent from the taxonomic origin². The FSPDs are rather intended to be the basis for analyses of sequence-structure-function relationships of protein families and to support protein engineering, than to reflect the organismic diversity. Thus, sequence redundancy due to deviating taxonomic sources is unessential. Using a user-defined threshold for sequence similarity within homologous families (default: 40%), all presorted protein entries are subsequently assigned to homologous families by grouping them to the most similar seed sequence or generation of additional homologous families³. Subsequently, the sequence entries are pushed into the database according to the pre-calculated organization in sequence and protein entries as well as the classification into different homologous families. For this purpose, the XML files belonging to the sequence entries are parsed in order to extract the protein names, eventually assigned EC numbers or annotations, taxonomic information and further accession numbers. All accession numbers and the respective source organism found for a sequence entry are stored into the database and linked to the respective amino acid sequence. Sorting as well as parsing and pushing are done on multiple CPU cores, if available.

Additionally, structures of proteins belonging to the respective protein family are integrated into the respective FSPD. Therefore, all proteins available in the Protein DataBank (PDB) (Berman et al. 2000) are compared to the sequences encompassed in the FSPD. In order to do this, all sequences of proteins with structure information provided by the PDB and a minimum length (default: 150 amino acids) are clustered into groups of homologous proteins with 50% identity.

²The information about different taxonomic origins of individual sequences that were combined to joined sequence entries are additionally deposited in the database.

³For further details on the classification see manuscript 'BioCatNet: a system for the analysis of sequence-structure-function relationships of protein families', [4](#) Section 4.4 on pages 127ff.

One sequence per group is defined as the representative and is compared to the local protein database. If the representative sequence of a cluster has an identity \geq a defined threshold (default: 40%) with any sequence already included in the FSPD, all structures of the respective cluster are analyzed concerning a potential family membership. Otherwise, all structures from the respective cluster are discarded. Thereby, the number of initial sequence comparisons is reduced by a factor of around 5 as compared to the number of entries in the PDB.

The subsequent family assignment during the integration of structures into the FSPD is based on the global sequence identity between the sequence of the respective protein structure and the most similar sequence in the database⁴. Identification of the most similar, available sequence is done using the fast *usearch_global* tool from the *usearch* program (Edgar 2010). According to the default values, structures identical with an sequence in the database are added to the respective sequence entry, structures with an identity of at least 98% are added to an existing protein entry by insertion of a new sequence entry, and structures with an identity of at least 40% are added to the database asking the curator for manual family assignment. Heteromultimeric protein structures encompassing multiple sequences of ThDP-dependent enzymes are linked to all contributing sequence entries in the FSPD. Notably, due to the extent of sequences in many protein families, it is beneficial for the performance to compare the clustered structures from the PDB against the FSPD rather than searching the previously collected sequences for structure information in the PDB.

Using *DBParse* instead of the previously used tools for database generation reduced the time-demand for the generation of a new version of the TEED enormously from several months to less than one week. *DBParse* is able to easily handle databases with an extent of ~ 170000 sequences (e.g. a FSPD of short-chain dehydrogenases⁵) and was further used for the generation of the most recent version of the TEED containing ~ 77000 sequences (Vogel and Pleiss 2014), an updated version of the CYPED containing ~ 19000 sequences⁶, an updated version of the LED encompassing ~ 80000 sequences⁷, an updated version of the LccED with ~ 10000 sequences⁸,

⁴Sequence identity is calculated by pairwise alignments *needle* from the EMBOSS package (Rice, Longden, and Bleasby 2000).

⁵Database generated and curated by Laurin Stahl and Silvia Fademrecht, Institute of Technical Biochemistry, University of Stuttgart, Germany

⁶Database generated and curated in collaboration with Łukasz Gricman, Institute of Technical Biochemistry, University of Stuttgart, Germany

⁷Database generated and curated by Silvia Fademrecht, Jennifer Häfner and Nora Schuth, Institute of Technical Biochemistry, University of Stuttgart, Germany

⁸Database generated and curated by Silvia Fademrecht, Institute of Technical Biochemistry, University of Stuttgart, Germany

the Imine Reductase Engineering Database (IRED) containing ~450 sequences⁸ (Scheller et al. 2014), an updated version of the LacED with ~500 sequences⁹, and an updated version of the TTCED encompassing ~800 sequences⁸.

2.1.3 DBUpdate

After the initial generation of a FSPD, manual curation of the family classification and enrichment by additional annotations, the database has to be updated in order to include sequences newly added to the NCBI protein database. In the course of updating, the established family classification and the existing annotations have to be preserved. Reusing parts of the *DBParse* algorithm and aiming for high performance and accuracy, Hannah Dienhart¹⁰ coded a program called *DBUpdate*, which can be used to incrementally update FSPDs generated by *DBParse*.

The general strategy of updating a FSPD is to search for homologous sequences that are not integrated yet and to insert those sequences into the corresponding homologous families. Thereby, homology searches could be performed using entire sequences or defined parts. *DBUpdate* allows updating FSPDs based on the global sequences of the encompassed proteins or by defined domains. Before updating a database, all sequences (of proteins or domains thereof) are clustered by *cluster_fast (usearch)* in order to identify centroids representing each cluster. These centroids are subsequently used to locally perform *BLAST* searches against the non-redundant protein database provided by the NCBI to identify homologous sequences. In order to avoid processing sequences already encompassed and sequences previously deleted from the database due to any reason, the obtained accession numbers of the identified homologs are compared to a list of accession numbers already processed during the database generation or previous updates. Only those accession numbers are further analyzed, which were never associated with the respective FSPD. Comparable to the approach used in *DBParse* all identified homologs are sorted into sequence and protein entries by taking both data sets, the existing database and the newly identified sequences, into account. Regardless of whether started from the entire sequences or specific domains for the homology search, the classification is generally done based on the global sequence identity. Subsequently, the newly found family members are added to the database. Accession numbers of sequences, which were identically or in extended form found in

⁹Database generated and curated in collaboration with Catharina Zeil and Silvia Fademrecht, Institute of Technical Biochemistry, University of Stuttgart, Germany

¹⁰*DBUpdate* was co-developed, programmed and tested by Hannah Dienhart in the context of her bachelor thesis.

the existing database, are added to the respective sequence entry. Newly found sequences are either added to existing proteins in case of a sequence identity of at least 98%, or inserted as new protein entries. Updating the experimentally determined structures can be done with the respective tool provided by *DBParse*.

2.1.4 BioCatNet

(For further details see manuscript 'BioCatNet: a system for the analysis of sequence-structure-function relationships of protein families', [\[4\]](#) Section 4.4 on pages 127ff.)

In order to allow for the systematic analysis of sequence-structure-function relationships, the database must provide detailed information about all three components, the sequence, the structure and the function of a multitude of diverse enzymes, in a computer-accessible format. As described, FSPDs implemented in the *DWARF* system can represent vast and diverse protein families like α/β -hydrolases, Cytochrome P450 monooxygenases and ThDP-dependent enzymes, including all publicly available sequences and structures. But the capability to additionally deposit functional information unambiguously associated with the sequence and structure information was lacking. In order to extend the applicability of FSPDs to actually serve as a basis for the systematic analysis of sequence-structure-function relationships, information on enzymes' functions has to be included into those databases. With *BioCatNet*, a successor of the *DWARF* system was developed, capable to encompass biochemical information in addition to sequence and structure information. The latter, information on sequences and structures of proteins, is publicly available, well organized, systematically searchable and easily accessible, whereas the majority of biochemical information derived from experiments is exclusively available in scientific literature. Distributed over a variety of different journals and books, documented with different accuracy and completeness, organized in tables, figures, supplementary information or just described in the text, this information is hard to access. Moreover, identification of the biocatalysts' sequences and thus unambiguous linkage of biochemical information and amino acid sequences often is challenging.

In order to make experimentally derived results available in databases, literature mining can be applied. However, due to the high complexity of scientific literature, the different possibilities to represent experimental results and the missing obligation to specify the amino acid sequence of the applied biocatalysts, the derived information has a good chance to be inadequately precise.

The alternative to literature mining is the direct data input of experimental data in electronic form into databases making these results easily accessible. With the increasing establishment of Laboratory Information Management Systems (LIMS), the classical lab journal is going to be replaced by an electronic alternative. LIMS help to make experimental data searchable but they are not implicitly intended to link function information with an amino acid sequence or even to make results publicly available. *BioCatNet*, a system combining a sequence and structure database, the capability to include experimental data linked to defined sequence entries, an intuitively useable GUI that can be used to insert data comparable to a LIMS and to provide information back to the experimenter, was designed and developed to overcome these drawbacks (Figure 4.13 on page 138). *BioCatNet* uses *DBParse* and *DBUpdate* (see Sections 2.1.2 and 2.1.3 on pages 23ff.) to incorporate sequence and structure information on specific protein families in FSPDs and provides a framework for the enrichment with biochemical information.

In collaboration with Michael Widmann¹¹, Jürgen Pleiss¹¹, Waldemar Reusch¹¹, Martina Pohl¹², Dörte Rother¹², Robert Westphal¹², Saskia Bock¹², Anna Baierl¹², Michael Müller¹³, Sabrina Loschonsky¹³, Maryam Beigi¹³ and Alexander Fries¹³ minimal standards for the acquisition of function information derived from experimental data were defined (Tables 4.6 and 4.7 on page 132). A data model fitting those requirements was developed based on the *DWARF* data model and optimized in collaboration with Waldemar Reusch¹⁴ (Figure 4.12 on page 137). The tables storing information on amino acid sequences, proteins and the family classification were designed based on the respective tables from the *DWARF* data model (Fischer 2004) but modified to correspond to the extended requirements. The additional tables were designed to enable storage of biochemical data and the taxonomic lineage.

Besides the technical challenge to design a relational data model capable to fulfill the mentioned requirements, a more practical question had to be answered: How to motivate bench scientists to actually insert their biochemical data into the TEED? The concept of *BioCatNet* was oriented towards data input by the experimenters rather than by literature mining in order to ensure maximal accuracy. This strategy certainly demands extra work from the experimenters since they have to make themselves familiar with the new system and they have to spend the time needed to insert the requested information into the respective forms. The answer to the asked

¹¹Institute of Technical Biochemistry, University of Stuttgart, Germany

¹²Institute of Bio- and Geosciences (IBG-1), Biotechnology, Forschungszentrum Jülich GmbH, Germany

¹³Department of Pharmaceutical and Medicinal Chemistry, University of Freiburg, Germany

¹⁴Diploma thesis, 2013-2014, Institute of Technical Biochemistry, University of Stuttgart, Germany

question is that data input into *BioCatNet* must be as simple and intuitive as possible to reduce the extra effort to a minimum and that the users have to receive a well-formated printout containing all their data that can be used to supplement classical lab journals. Moreover, the graphical user interface (GUI) should allow access to all released information on sequences, structures and functions enriched by helpful information in order to support the scientists' work. Since *BioCatNet* and the associated software tools further were intended to replace the *DWARF* system, the GUI had to provide all the relevant features needed to maintain, analyze and publish FSPDs. Started as a research assistant and continued in the context of his diploma thesis, Waldemar Reusch developed a web-accessible GUI meeting those requirements (www.biocatnet.de).

2.2 Sequence

Approaching the first aspect of the Sequence-Structure-Function Relationships of ThDP-dependent enzymes

FSPDs enable the identification of functionally or structurally relevant positions based on the amino acid composition within protein families. This is based on the assumption, that the mutational pressure on different positions in proteins is affected by their functional or structural relevance. Positions needed for catalytic activity, substrate recognition, activation, correct folding of the primary structure into tertiary structure, assembly of oligomeric structures, or stabilization of the protein are more likely to keep defined amino acids at those positions during the course of evolution as compared to less relevant positions. Such positions can be identified by analyzing multisequence alignments of protein families for conservation of distinct positions or correlated mutational behavior of several positions. Thus, multisequence alignments (MSA) are a prerequisite for this type of sequence analysis. MSAs in their classical form represent biologically comparable positions by arranging them in columns. By insertion of gaps, variations in sequence lengths due to inserts and deletions are illustrated.

In order to enable position-specific annotation and analysis, the *DWARF* database system was designed to store each amino acid of any protein as a single entity (Fischer, Thai, et al. 2006). As a consequence, for each analysis

- a set of amino acid sequences had to be generated by concatenation of single positions,
- the resulting sequences had to be used for generation of multisequence alignments,
- the alignments had to be investigated for presumable functionally relevant positions,
- and the resulting information had to be inserted into the database as annotations.

Thus, since comparison of positions from different proteins depended on intermediate alignment generation, the *DWARF* system did not support position-specific analysis in a straightforward manner. Although each single position was uniquely addressable by using an ID, lack of implicit information on the comparability of different positions from different sequences hindered analysis. In this work, a method for assigning family-specific standard numbers to all positions of homologous sequences based on structure information was developed. Application of this method subsequently allowed to address each position and all other structurally equivalent positions directly from the protein database without the need for intermediate alignment steps.

2.2.1 A standard numbering scheme for ThDP-dependent decarboxylases

(For further details see publication 'A standard numbering scheme for thiamine diphosphate-dependent decarboxylases', [DOI](#) Section 4.1 on pages 76ff.)

Global alignments of sequences belonging to one protein family, regardless of whether pairwise or with multiple sequences, aim to represent the equivalence of the C α -atoms. Thus, positions aligned by multisequence alignments are intended to be 'structurally equivalent'. Consequently, incorporation of structure information during the alignment process was shown to improve the alignment quality (Notredame, Higgins, and Heringa 2000). In order to analyze the amino acid sequences of protein families for conservation or correlated mutations, it would be desirable to be able to address the structurally equivalent positions of different proteins with a common identifier. To provide a method, which on the one hand has a high chance that the addressed positions are structurally equivalent and on the other hand can be applied to high numbers of proteins, an algorithm incorporating structure alignments and sequence-profile alignments was developed. Based on a structure alignment of representative ThDP-dependent DCs generated by a modified¹⁵ version of STAMP (Russell and Barton 1992), a hidden Markov model (HMM) was created using HMMER (*HMMER*, <http://hmmmer.janelia.org/> 2013). The HMM profile could subsequently be used to align any query sequence belonging to the same protein family against a predefined reference sequence. By transferring the absolute position numbers of the reference sequence to the aligned positions of the query sequence, standard position numbers were assigned (Figure 2.1 on the next page). The pyruvate decarboxylase from *Saccharomyces cerevisiae* (ScPDC) was chosen as the reference sequence for the standard numbering scheme for ThDP-dependent DCs. A similar approach was previously chosen for the standard numbering scheme for metallo- β -lactamases (Ambler et al. 1991; Galleni et al. 2001; Garau et al. 2004), but not a single sequence was defined to serve as the reference but an alignment of multiple family members. Using one reference sequence instead of a reference alignment as the source of the standard numbers

¹⁵Modifications to *STAMP* v4.4 were necessary in order to increase the number of alignable structures due to memory usage restrictions in the original version.

alignfit.h:	lines 42-43 changed to	#define MAXslen 100000
		#define MAXnbloc 10000
	line 44 changed to	#define MAXtlen 2000
alignfit.c:	line 95 changed to	parms[0].MAX_SEQ_LEN=100000;
poststamp.c:	line 41 changed to	#define MAX_SEQ_LEN 100000

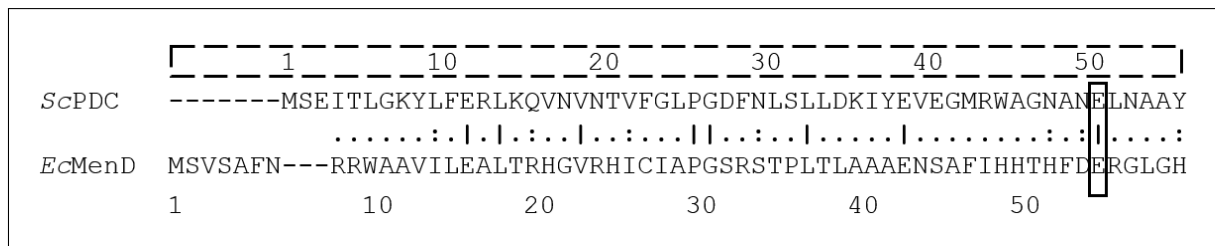


Figure 2.1: Schematic representation of the standard number assignment. After a profile-guided alignment of a query sequence (here: 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexadiene-1-carboxylate synthase (MenD) from *Escherichia coli*, *EcMenD*) and the reference sequence (pyruvate decarboxylase from *Saccharomyces cerevisiae*, *ScPDC*), the absolute position numbers of the reference sequence (dashed) were transferred to the respective positions of the query sequence. The cofactor activating glutamate at position 55 in *EcMenD* could henceforth be addressed using the standard position number 51.

during the number assignment confers robustness against expected changes in the sequence space. With the increasing number of known amino acid sequences belonging to a protein family, modification of a reference alignment becomes necessary in order to assure continuous conformance with the entire protein family. As a consequence, standard numbers once assigned to positions might change over time, as Dr. Michael Widmann¹⁶ observed during his work with metallo- β -lactamases (Widmann, Pleiss, and Oelschlaeger 2012). Based on this observation, the described method for standard number assignment was developed intended to be robust against changes of the known sequence space over time. In case of using a reference sequence, even if updates of the HMM profile used for the alignment generation might become necessary, changes of the sequence space do not affect the standard numbers once assigned to positions of any sequence in the respective protein family. By making use of the family-specific HMM profile during the alignment of the query and the reference sequences, the method simultaneously benefits from the increased alignment accuracy of multisequence alignments as compared to pairwise alignments.

In the course of developing the standard numbering method, a specific file format (called 'nvw'¹⁷) was established, which allows to store the results of a standard number assignment for any sequence. In addition to the amino acid sequence and the standard numbers assigned to all positions, 'nvw' files may contain position-specific annotation information. The program developed for the standard number assignment is able to incorporate position-specific annotation information as defined by 'anno' files (Listing 2.1 on page 35) and to transfer the defined

¹⁶Institute of Technical Biochemistry, University of Stuttgart, Germany

¹⁷The 'nvw' (numbering Vogel-Widmann) file format was established in order to allow storage of standard numbers once assigned to amino acid sequences. For detailed file format declaration see Supplementary Information, Listing A.1 on page 194

annotation information to the respective residues of the processed amino acid sequence.

Application of the standard numbering scheme to all sequences of the DC superfamily of the most recent version of the TEED (version 11.5) and subsequent analysis of the amino acid distribution revealed 23 positions to be conserved in more than 80% of those sequences (Table 2.1 on page 36). For a majority of those conserved positions, a functional or structural role was not experimentally shown. Standard position 51 carries the glutamic acid residue responsible for the activation of the ThDP cofactor, whose binding is facilitated by the conserved positions 443, 444 and 471 of the GDGX_{24,27}N-motif (Hawkins, Borges, and Perham 1989). Conserved glycine residues are located at the C-cap of α -helices at standard positions 27, 75 and 219, at the N-cap of an α -helix at standard position 416 and on the hinges of loops at standard positions 161, 286, 443 and 506. In addition, two further conserved glycine residues were identified at standard positions 77 and 424¹⁸. Position 77 forms part of the active site cavity near the pyrimidine moiety of the ThDP cofactor and position 424 is part of an α -helix in the core of the PP domain. Besides the N-terminal methionine, which is in most ThDP-dependent DCs referred to as standard position 0.1, nine further conserved positions with yet unknown functional or structural role were identified. In order to allow application of the standard numbering scheme to any ThDP-dependent DC, a web application was implemented in the *BioCatNet* system housing the TEED¹⁹.

Multisequence alignments based on the standard numbering scheme for ThDP-dependent decarboxylases

Furthermore, an alignment program called *nvwAlign* was developed, which allows to align multiple 'nvw' files to a multisequence alignment. By reading multiple 'nvw' files, virtual formation of an two-dimensional array with the columns reflecting the standard numbers in ascending order and the rows corresponding to the multiple sequences, and subsequent output of this array as an alignment, *nvwAlign* allows to align multiple sequences. Since the standard numbers read from the 'nvw' files were assigned using an HMM profile, which in turn was derived from a structural alignment, the generated multisequence alignment represents the structural equivalence of amino acids arranged within alignment columns. The program allows

¹⁸For more details about conserved glycine residues in ThDP-dependent enzymes see the publications in [\[1\]](#) section 4.1 on page 76 and [\[2\]](#) section 4.2 on page 92.

¹⁹www.teed.biocatnet.de/workbench/numbering

Listing 2.1: Annotation ('anno') file used during the standard number assignment of ThDP-dependent DCs to transfer annotation information to the respective positions of the query sequence. The containing information on positions, corresponding annotation description and color for highlighting the respective positions in an alignment are transferred to the 'nvw' file. The '>' and '<' are used to dynamically determine the start and end positions of annotated regions. Thus, the borders of region-annotations are automatically adjusted up- or downstream in case of gaps at the respective positions. E.g. the region between standard positions 6 and 168 is going to be annotated as the PYR domain of ThDP-dependent enzymes.

```
# 2014, ITB Bioinformatics , Constantin Vogel
#
# - copy this file to "annotation_[yourproject].anno and edit the entries
# - store the new file into "aln_hmm"
# - edit $annotation_file in donumbering_[yourproject].pl
#
# scheme:
# residue-position | annotation-description | annotation-color
#
# example:
# 20 | substrate recognition/binding | #CC6600
# 25 | substrate recognition/binding | #CC6600
#
# reference 2vk8
#
6> | PYR-start | #FC0C02
168< | PYR-end | #FC0C02
197> | TH3-start | #FF8C00
336< | TH3-end | #FF8C00
367> | PP-start | #053BF9
540< | PP-end | #053BF9
25 | S-pocket | #ffdead
26 | S-pocket | #ffdead
27 | S-pocket | #ffdead
28 | S-pocket | #ffdead
51 | ThDP-binding | #9932cc
114 | donor-pocket | #ffdead
115 | donor-pocket | #ffdead
166 | PX motif | #uu6785
167 | PX motif | #uu6785
168 | PX motif | #uu6785
169 | PX motif | #uu6785
170 | PX motif | #uu6785
171 | PX motif | #uu6785
221 | activator binding | #ffdead
443 | GDGX-motiv | #FFCCFF
444 | GDGX-motiv, Mg+ binding | #CC6600
445 | GDGX-motiv | #FFCCFF
446 | GDGX-motiv, TPP-binding | #00FF33
471 | Mg+ binding | #CC6600
473 | Mg+ binding | #CC6600
476 | entrance to S-pocket | #ffdead
477 | S-pocket | #ffdead
```

Table 2.1: Conserved positions ($\geq 80\%$) in DCs identified using the standard numbering scheme for ThDP-dependent DCs.^[a]

std. pos.	0.1	14	27 ^[b]	51 ^[b/c]	58 ^[b]	75	77	88
	M 89%	L 83%	G 90%	E 94%	A 97%	G 91%	G 83%	A 88%
		F 1%	S 2%	V 3%		A 4%	A 15%	S 6%
								C 3%
std. pos.	94 ^[b]	152	161	168	219 ^[b]	241	280 ^[b]	286 ^[b]
	P 92%	A 88%	G 84%	P 87%	G 93%	P 86%	D 90%	G 97%
	K 3%	V 5%	R 7%	A 5%	D 5%	G 3%	E 3%	D 1%
	A 1%	C 1%	K 2%	T 2%		A 3%	Q 3%	
	G 1%	G 1%	A 2%	S 2%		R 2%		
		L 1%	Q 1%	C 1%		V 1%		
						L 1%		
std. pos.	416	422	424 ^[c]	443 ^[b/c]	444 ^[b/c]	471 ^[c]	506	
	G 83%	A 80%	G 83%	G 98%	D 95%	N 85%	G 82%	
	D 10%	G 7%	A 14%		E 4%	D 13%	D 5%	
	A 6%	S 4%	S 2%				N 3%	
		T 3%					E 2%	
		L 2%					R 1%	
		I 1%					Q 1%	
		C 1%					H 1%	
		V 1%						

^[a] Only amino acids occurring at the respective positions in at least 1% of all analyzed sequences are shown.

^[b] Position has previously been described to be conserved in more than 90% of all DCs (Vogel, Widmann, et al. 2012, see [1](#) Section 4.1.3 on pages 84ff.).


^[c] Position has previously been described to be conserved in more than 80% of all ThDP-dependent enzymes (Vogel and Pleiss 2014, see [2](#) Section 4.2.4 on pages 105ff.).

Standard positions 60 and 445, which carry glycine in 82% and 80% of all ThDP-dependent enzymes (Vogel and Pleiss 2014, see [2](#) Section 4.2.4 on pages 105ff.), respectively, were found to be less conserved within the superfamily of DCs (65% and 75% glycine, respectively).

to output the alignment either in the CLUSTAL alignment format or as an HTML file, both either in blocks of defined length or in an unwrapped manner. The CLUSTAL format has the advantages of being compatible with a multitude of bioinformatics tools and its convertibility to other alignment formats while the proprietary alignment in HTML format provides more comfort. Visualized by any web browser application, the generated alignments in HTML format provide annotations highlighted by the colors defined in the 'anno' file (Listing 2.1 on the previous page) and sequence specific, absolute position numbers plus the assigned standard position numbers for all amino acids in the alignment while hovering the respective amino acids with the mouse cursor.

Kindly supported by Dr. Michael Widmann, the standard numbering method was validated by comparing the standard numbers assigned to sequences with known crystal structures with the structural equivalence of those positions between the query and the reference sequences. In addition, *nvwAlign* was compared to the alignment program *T-Coffee* (Notredame, Higgins, and Heringa 2000), where it outperformed *T-Coffee* in the alignment quality and matched its calculation performance (for the complete validation alignment see Supplementary Information, Section A.4.3 on pages 197ff.).

2.2.2 A standard numbering scheme for the PYR and PP domains of ThDP-dependent enzymes

(For further details see publication 'The modular structure of ThDP-dependent enzymes',  Section 4.2 on pages 92ff.)

Although the ThDP-dependent DCs are the most comprehensive superfamily of the TEED containing 30% of all known ThDP-dependent enzymes, the limitations of a standard numbering scheme restricted to DCs are obvious. In order to allow for the systematic analysis of sequences from different superfamilies and to be able to transfer knowledge about functionally relevant positions from the DC superfamily to others, a standard numbering scheme capable to be applied to the sequence of any ThDP-dependent enzyme was desirable. Direct application of the standard numbering scheme developed for the ThDP-dependent DCs to other superfamilies by expansion of the profile HMM was not possible due to different domain architectures on the sequences of different ThDP-dependent enzymes (Figure 2.2 on the following page). However, due to sequence and structure similarity of the two catalytic domains (PYR and PP domain), generation of profile HMMs of those domains was possible. Based on a superimposition of 50 and 48 structures of PYR and PP domains, respectively, a multisequence alignment was derived, expanded by additional sequences from homologous families without available structure information, and was used for the generation of two separate profile HMMs.

In order to define standard numbers, the *ScPDC* sequence, which was used as the reference sequence for the standard numbering scheme for the DC superfamily, was aligned with the profiles of the PYR and PP domains and the 'match' states of both profiles were named according to the aligned positions of the reference sequence. Subsequently, both profiles were aligned with the respective best matching parts of the sequences of ThDP-dependent enzymes using

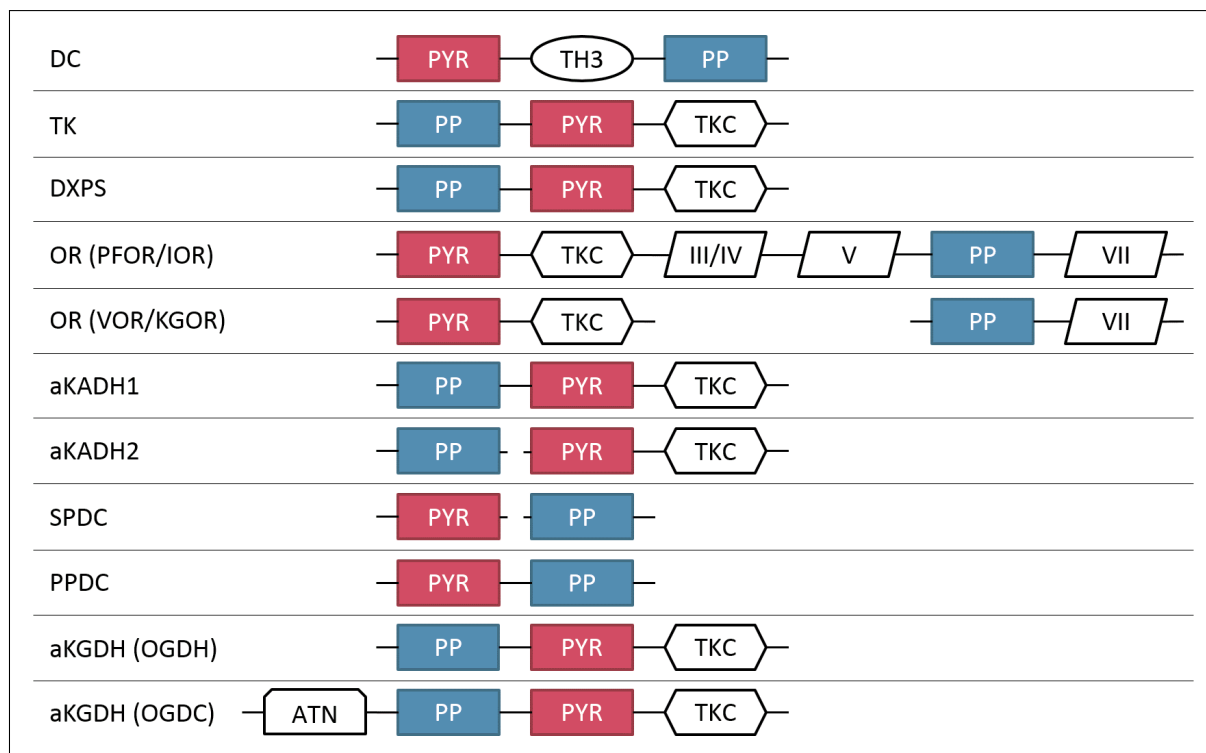


Figure 2.2: Order of different domains on the primary structure of ThDP-dependent enzymes (derived from Widmann, Radloff, and Pleiss 2010). The different superfamilies of ThDP-dependent enzymes differ in their domain composition and the order of the different domains on their sequences. Although all ThDP-dependent enzymes possess a PYR and a PP domain, the occurrence of additional domains varies. In ORs, the last domain (VII or γ) can also be found as a separate gene product. (DC: decarboxylases, TK: transketolases, DXPS: 1-Deoxy-D-xylulose-5-phosphate synthases, OR: oxidoreductases, PFOR: pyruvate:ferredoxin oxidoreductases, IOR: indolepyruvate:ferredoxin oxidoreductases, VOR: 2-ketoisovalerate:ferredoxin oxidoreductases, KGOR: 2-ketoglutarate:ferredoxin oxidoreductases, aKADH1: α -ketoacid dehydrogenase superfamily 1, aKADH2: α -ketoacid dehydrogenase superfamily 2, SPDC: sulfopyruvate decarboxylases, PPDC: phosphonopyruvate decarboxylases, aKGDH: α -ketoglutarate dehydrogenase superfamily, OGDH: 2-oxoglutarate dehydrogenases, OGDC: 2-oxoglutarate decarboxylases, PYR: pyrimidine-binding domain, PP: pyrophosphate-binding domain, TH3: transhydrogenase domain dIII, TKC: transketolase C-terminal domain, ATN: N-terminal acyltransferase-like domain).

hmmsearch from the *HMMER* toolbox (*HMMER*, <http://hmmmer.janelia.org/> 2013). By transfer of the previously assigned standard position numbers from the 'match' states to the respectively aligned positions of the query sequences, the PYR and PP domains of all sequences in the TEED were numbered. Subsequently, the entire family was analyzed for highly conserved positions with one specific amino acid in more than 80% of all sequences. Seven such positions were identified including the GDGX_{24,27}N-motif (Hawkins, Borges, and Perham 1989), the cofactor activating glutamate, described to be conserved in the majority of all ThDP-dependent enzymes (Shaanan and Chipman 2009; Candy, Koga, et al. 1996; Lee, Lee, et al. 2013), and two glycine residues (Figure 4.5 on page 105)²⁰. Notably, with the exception of standard position 60, those

²⁰More details are described in the publication 'The modular structure of ThDP-dependent enzymes' on pages 92ff.

positions were unsurprisingly also found to be conserved in the DC superfamily (Table 2.1 on page 36).

Moreover, the standard numbering scheme for the PYR and PP domains of ThDP-dependent enzymes enabled a compilation of literature about variants organized by the distinct positions. Therefore, names of the different superfamilies and homologous families of ThDP-dependent enzymes were used as queries for a literature search in PubMed (www.ncbi.nlm.nih.gov/pubmed). An extensive examination of the identified publications²¹ for mentioned enzyme variants²², manual assignment of actual amino acid sequences to each publication and identification of the standard numbers of each of the found variants enabled comparison of corresponding mutations in different enzymes (Supporting Information, Table A.1 on pages 177ff.).

²¹The list of publications was extended by articles that were referred to in the identified publications, by the collection of related literature that was previously known to the author and by kind support of Prof. Dr. Martina Pohl, who provided further literature.

²²The examination of publications' abstracts and full texts was kindly supported by Łukasz Gricman, who provided a tool for the identification of enzyme variants on literature.

2.3 Structure

Approaching the second aspect of the Sequence-Structure-Function Relationships of ThDP-dependent enzymes

2.3.1 The modular structure of ThDP-dependent enzymes

(For further details see publication 'The modular structure of ThDP-dependent enzymes', [\[2\]](#) Section 4.2 on pages 92ff.)

For the generation of the domain-based standard numbering scheme for all ThDP-dependent enzymes (Section 2.2.2 on pages 37ff.), all available structures of family members were aligned relative to their catalytically relevant PYR and PP domains. Based on this alignment and the relative position of the domain boundaries in the structure of *ScPDC* (pdb|2VK8, Kutter et al. 2009), the borders of the PYR and PP domains were identified, if they exist (Supplementary Information, Table A.6 on pages 213ff.). Structure information for ThDP-dependent enzymes is available for proteins from seven of nine superfamilies. While no experimentally determined structure was available for the sulfopyruvate decarboxylases (SPDC) and phosphonopyruvate decarboxylases (PPDC), multiple structures were published for DCs, TKs, DXPSs, oxidoreductases (OR), α -ketoglutarate dehydrogenases (aKGDH) and both superfamilies of α -ketoacid dehydrogenases (aKADH1 and aKADH2, previously also called K1 and K2). The superimposition of the available structures furthermore revealed a conserved structural orientation of two pairs of PYR and PP domains, forming two active sites. However, the structures of members of different superfamilies deviate considerably concerning the additional domains, like the TH3 and TKC domains, and the loops connecting the more conserved PYR and PP domains. In consideration of the deviating composition by different domains and the deviating arrangement of those domains on the sequences of different superfamilies (Figure 2.2 on page 38), structural accordance of the PYR and PP domains implies different 'wiring' of those domains. Moreover, comparison of the TK and DXPS structures, with identical sequential domain arrangement and higher sequence similarity of both superfamilies as compared to others, also revealed different structural architectures. Detailed investigation brought up six different structural architectures ([\[2\]](#) Figure 4.4 on page 101) belonging to five different basic layouts to represent all known structures of ThDP-dependent enzymes ([\[2\]](#) Figure 4.3 on page 100). By distinction between 'PYR-PP', 'PP-PYR' and 'uncoupled types', depending on the linkage and the relative order on the sequences, and differentiation of the active site composition in 'intra-monomer' or 'inter-

monomer' types, the architectures of ThDP-dependent enzymes were described. 'Intra-monomer' describes active sites formed by a pair of PYR and PP domains belonging to the same monomer, whereas 'inter-monomeric' active sites are constituted by the interplay of PYR and PP domains of different monomers. In addition, an evolutionary pathway from a putative homomultimeric enzyme consisting of only one domain to the different contemporary ThDP-dependent enzymes was drawn, combining global sequence similarities, similarities of the more conserved PYR and PP domains, and the different structural architectures (Figure 4.8 on page 115).

2.3.2 Structural rearrangement of the *ApPDC*

Based on the findings about different arrangements on the sequences and structures of ThDP-dependent enzymes, a variant of the *ApPDC* with altered domain-order was designed. Since the core structure consisting of two pairs of PYR and PP domains forming two active sites remains unaffected by the observed architectural differences, a variant was designed to investigate, which influence the different architectures have on the respective function of enzymes from different superfamilies. *ApPDC* belongs to the DC superfamily and the wildtype (wt) structure (pdb|2VBI, Rother et al. 2011) reflects the typical inter-monomer/PYR-PP architecture with a N-terminal PYR domain and a C-terminal PP domain linked by an additional TH3 domain. The two active sites of the homodimeric complex are formed at the interface between two monomers, so that both monomers concertedly participate in the binding of the cofactor ThDP: at the pyrophosphate part by the PP domain of one monomer and at the pyrimidine part of the cofactor by the PYR domain of the other monomer. The variant was designed to form an intra-monomer/PP-PYR type structure, which is the opposite of the inter-monomer/PYR-PP architecture of the wildtype enzyme. An intra-monomer/PP-PYR architecture was observed for DXPSs, which in contrast to the DCs obtain a C-terminal TKC domain and are lacking the DC-typical TH3 domain. Because it was shown that the TKC domain is not essential for catalytic activity in the transketolase of *Escherichia coli* (Costelloe, Ward, and Dalby 2008) and that the TH3 domain of the DC superfamily has lost its nucleotide-binding role in many DCs (Duggleby 2006), the variant was designed without additional domains besides the PYR and PP domains to minimize the complexity. Although detailed structure information is missing for PPDC and SPDC, gel filtration chromatography and activity screenings suggested multimeric complexes of these proteins to form the active enzymes (Graupner, Xu, and White 2000; Johnen and Sprenger 2009). Moreover,

since both families are lacking additional domains besides the PYR and PP domains, it stands to reason that the additional domains rather play a structural than a functional role. Consequently, omission of the TH3 domain was supposed not to influence the activity. Thus, the PP-PYR variant of *ApPDC* represents a modified, circular-permutation of the wt*ApPDC*. In order to link the N-terminal PP to the C-terminal PYR domain, the C-terminal helix of the PP domain was modified to allow coupling to the N-terminus of the PYR domain.

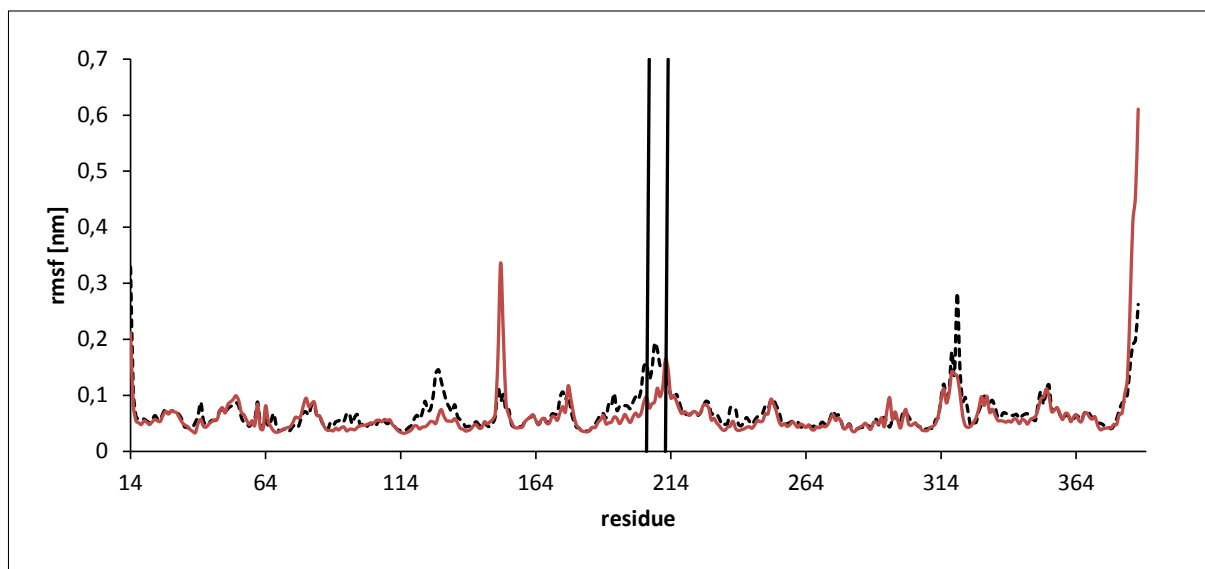


Figure 2.3: Root mean square fluctuation of the intra-monomer/PP-PYR type *ApPDC* variant. Both monomers forming the active dimer are shown (chain A, dashed black line; chain B, red solid line). By circular permutation, optimization of the linker connecting the former N- and C-termini and removal of the TH3 domain, a new variant of *ApPDC* with altered sequence and structure architecture was designed. MD simulation for 10 ns and analysis of the RMSF revealed moderate flexibility of the designed linker and two further flexible regions at the protein surface. The linker region is shown framed by two vertical lines.

Three different linkers were designed and modeled using the program 'Yasara' (Krieger, Koraïmann, and Vriend 2002). Subsequently, the protein variants were simulated for 10 ns with the Molecular Dynamics (MD) software 'Gromacs' (Spoel et al. 2005) in order to analyze the stability of three different variants. The simulations were done with the AMBER03 force field (Duan et al. 2003) in SPC/E water (Berendsen, Grigera, and Straatsma 1987), including one cofactor ThDP and a Mg^{2+} ion per active site plus Na^+ counter ions to compensate the negative charge of the proteins. Sven Benson²³ kindly provided the force field for the ThDP cofactor. Analysis of the root mean square fluctuation (RMSF) of the $C\alpha$ atoms in the protein backbone revealed moderate flexibility of the linkers and two residues at the protein surface (Figure 2.3).

²³Sven Benson, Institute of Technical Biochemistry, University of Stuttgart, Germany

The variant²⁴ with the lowest linker flexibility was chosen for experimental investigation of expressibility and activity. The designed enzyme was purchased as a synthetic gene from GeneArt[®] Gene Synthesis (life technologies[™]). Cloning into *E. coli* BL21 DE3, expression and analysis of the biochemical properties was kindly performed by Saskia Bock²⁵. The calculated mass of the designed enzyme construct was 40.6 kDa and a SDS polyacrylamide gel electrophoresis of the crude extract of transfected cells indeed revealed soluble protein (Figure 2.4). Although weakly over-expressed, soluble protein with a mass of ~ 41 kDa was observed in the crude cell extract, successful expression of the designed enzyme variant could not be approved, since the same band was observed after 24h in a negative control with an empty vector.

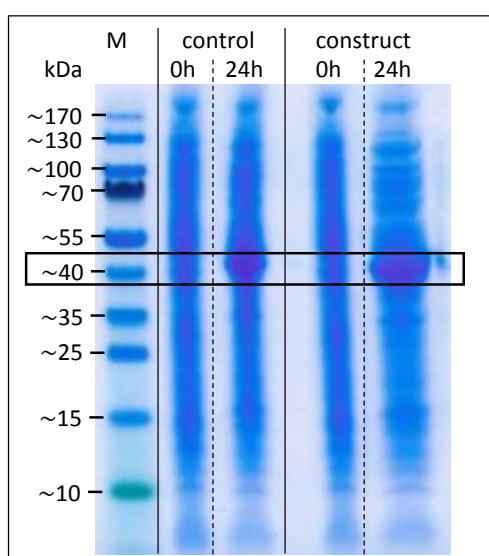


Figure 2.4: SDS gel of the crude extract of cells with and without the synthetic gene for a designed *ApPDC* variant kindly provided by Saskia Bock. Over-expressed, soluble protein with a mass off ~40-43 kDa was detected for the negative control and the transfected cells (black horizontal box). M: marker; control: crude extract of *E. coli* BL21 DE3 cells after transformation with an empty vector as negative control; construct: crude extract of *E. coli* BL21 DE3 cells after transformation with a vector carrying the synthetic gene for the designed *ApPDC* variant.

2.3.3 Automated homology modeling

The recent version of the TEED contains 240 crystal structures. Out of the 77493 sequences in the TEED, 284 (0.04%) were found to be part of those crystal structures. In order to facilitate analysis of the sequence-structure relationship of ThDP-dependent enzymes, enlargement of this fraction was desirable. Thus, a workflow for the automated generation of homology models was developed. Each sequence without available structure information (referred to as 'target') was compared to all available structures serving as templates. The sequentially most similar template for each target was chosen for modeling. In order to increase the probability of accurate modeling,

²⁴The variant consists of additional amino acids (shown in quotation marks) and wt*ApPDC* residues (defined by the position numbers as found in the structure pdb|2VBI): 359-543, 546-552, 'SG', 555-560, 5-174 and 'IEV'. For the complete amino acid and DNA sequences, see Supplementary Information, Section A.1 on page 174.

²⁵Institute of Bio- and Geosciences (IBG-1), Biotechnology, Forschungszentrum Jülich GmbH, Germany

only sequences with a global sequence similarity $>60\%$ ²⁶ to the respectively selected template were modeled. The target-template alignments and the homology models were generated using *MODELLER* 9.13 (Webb and Sali 2014). For each target, ten models were predicted in parallel and based on the objective function implemented in *MODELLER* (Sali and Blundell 1993), the best two models were selected for inclusion in the TEED. Using all ten generated models per target sequence, the root mean square deviation (RMSD) of the $C\alpha$ atoms was calculated and plotted. In addition, DOPE (Discrete Optimized Protein Energy) scores were calculated for the two best models and plotted superimposed with the DOPE scores for the corresponding positions of the template (Figure 2.5 on the following page) (Shen and Sali 2006). The plots of the DOPE scores and the RMSDs of independently generated models can be used to evaluate the model quality. DOPE profiles, which deviate between the target and template, reveal wrong sequence alignments, which in turn results in inaccurate models. The RMSD of the $C\alpha$ atoms of different predicted models shows regions of low structural determination, which correspond to regions considerably deviating between the target and the template as well as regions with missing structure information in the template. 52646 models of 26323 sequences were generated. Finally, in order to validate the predicted structures, the location of the conserved GDGX motif of ThDP-dependent enzymes was compared between all models and the respective templates. Therefore, the GDGX motif was identified in the sequences of all PP domains using the domain specific HMM profile (for details on the domain specific HMM profiles see Section 2.2.2 on page 37). Using the coordinates of the $C\alpha$ atoms of the GDGX motif, a local RMSD between each model and the respective template was calculated. For 87%, 3% and 10% of all models, the RMSD was less than 1 Å, between 1 Å and 4 Å or >4 Å, respectively. Visual inspection of a random sample of models with an RMSD between 1 and 4 Å revealed slight shifts between the models and templates in the structural superimposition but correct modeling. Consequently, only 10% of the models were removed due to alignment and subsequent modeling errors. For all remaining models, probable multimeric structures were predicted based on the oligomeric state of the respective template. Using *PyMol* (*The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC, <http://pymol.org> 2013*) for performing structure duplication, protein chain renaming and structure alignments, homo- and heteromultimers were constructed. Generation of homomultimers required duplication of the modeled monomer and alignment of both

²⁶The global similarity was calculated by pairwise target-template alignments using the *needle* tool from the EMBOSS software suite (Rice, Longden, and Bleasby 2000).

monomers with the two monomers of the respective template structure. For heteromultimeric enzymes, monomers modeled for sequences with the same source organism were combined if they were initially modeled against the same template structure. Thus, 19777 multimeric models were generated for ThDP-dependent enzymes lacking experimentally determined structure information.

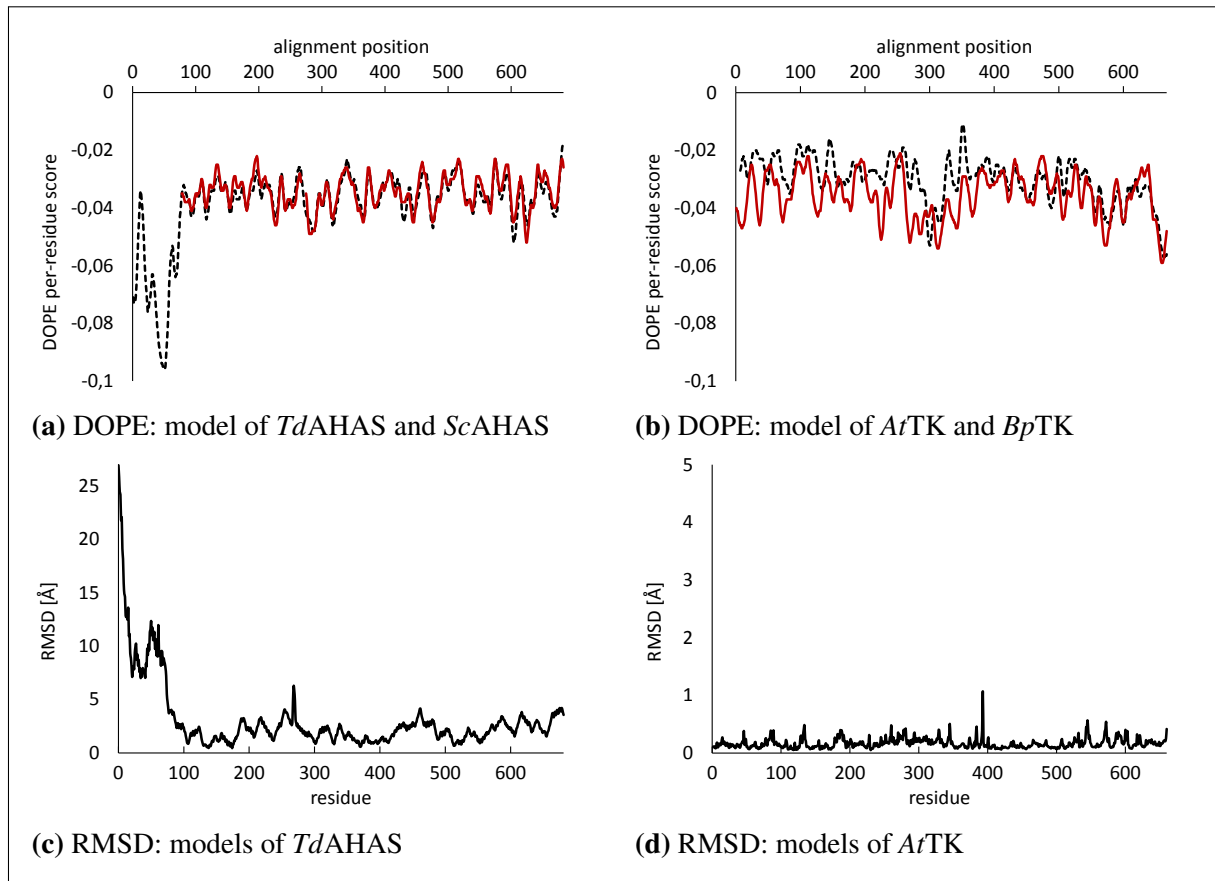


Figure 2.5: The DOPE scores and RMSDs of homology models of two sequences from the TEED against two structure templates allow to evaluate the model quality. A putative acetohydroxyacid synthase (AHAS) from the yeast *Torulaspora delbrueckii* (*TdAHAS*, sid|11616) and a transketolase from *Agrobacterium tumefaciens* (*AtTK*, sid|29832), both missing experimentally determined structure information, were modeled using the structures of *ScAHAS* (pdb|1N0H, Pang et al. 2004) and the transketolase from *Burkholderia pseudomallei* (*BpTK*, pdb|3UPT, Baugh et al. 2013) as templates, respectively. The DOPE scores (a/b) were calculated using *MODELLER* 9.13 (Webb and Sali 2014), the RMSDs (c/d) were calculated from the three-dimensional coordinates from ten individual models. (a) The DOPE scores calculated for a homology model of the *TdAHAS* sequence (black, dashed line) highly resemble the DOPE scores calculated for the used template structure of *ScAHAS* (red, solid line), confirming accurate alignment of both sequences. Due to missing structure information in the template for the N-terminus in the alignment, the respective region in the modeled structure is not reliable. (b) Significant differences in the DOPE scores calculated for the model of *AtTK* (black, dashed line) and the template *BpTK* (red, solid line) reveal a defective target-template alignment, resulting in a corrupt homology model. The respective model was automatically recognized and discarded. (c) The RMSD of ten individual models of the *TdAHAS* sequence using the same template (pdb|1N0H) shows regions of unreliable quality. As a consequence of the missing structure information in the template for the N-terminus, the different models vary in the respective region, emphasizing arbitrary modeling of the N-terminus. (d) Low RMSD of the models generated for the *AtTK* sequence indicate constant modeling quality in all ten modeling runs. However, due to the incorrect target-template alignment, the models are meaningless.

2.4 Function

Approaching the third aspect of the Sequence-Structure-Function Relationships of ThDP-dependent enzymes

For all subsequently mentioned amino acid positions, the respective standard position numbers are given in parenthesis. For details on the standard number assignment see section 2.2.1 on pages 32ff.

2.4.1 Rational engineering of a ThDP-dependent enzyme for the direct asymmetric synthesis of (*S*)-benzoin

ThDP-dependent enzymes proved to be able to catalyze a broad range of reactions including the enantioselective carbonylation of aldehydes, α -keto acids and ketones into chiral α -hydroxy ketones (Hoyos et al. 2010; Brovetto et al. 2011). However, enzymatic, direct asymmetric synthesis of (*S*)-benzoin by homocoupling of benzaldehyde and benzaldehyde derivatives has not been achieved so far. Thus, synthesis of (*S*)-benzoin was limited to enzymatic kinetic resolution and deracemization of racemic benzoin (Demir, Pohl, et al. 2001; Fragnello et al. 2012), asymmetric reduction of benzils (Demir, Peruze, et al. 2008) or chemical synthesis (Enders and Kallfass 2002b; Enders and Kallfass 2002a). Engineering of the *Ap*PDC previously provided access to (*S*)-phenylacetylcarbinol (PAC) by opening the *S*-pocket (Rother et al. 2011). According to the *S*-pocket concept (Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008), substitution of the glutamic acid at position 469 in *Ap*PDC (477) by glycine, allowed anti-parallel arrangement of the cofactor-bound donor substrate acetaldehyde and the acceptor substrate benzaldehyde. While the wildtype shows the DC-typical (*R*)-selective behavior in the carbonylation of benzaldehyde and acetaldehyde, the *Ap*PDC-E469G variant allowed synthesis of (*S*)-PAC. However, synthesis of (*S*)-benzoin was prevented by the size of the donor-binding site, which did not allow binding of benzaldehyde (Figure 2.6 on the next page). A larger donor-binding site capable to bind benzaldehyde was described for the benzaldehyde lyase from *Pseudomonas fluorescens* (*Pf*BAL), which was shown to be a highly efficient but (*R*)-specific catalyst for the homocoupling of benzaldehyde to benzoin (Demir, Sesenoglu, Eren, et al. 2002; Janzen et al. 2006). According to the *S*-pocket concept, (*R*)-specificity of *Pf*BAL could be explained by the insufficiently large acceptor-binding site, preventing antiparallel arrangement of two benzaldehyde molecules.

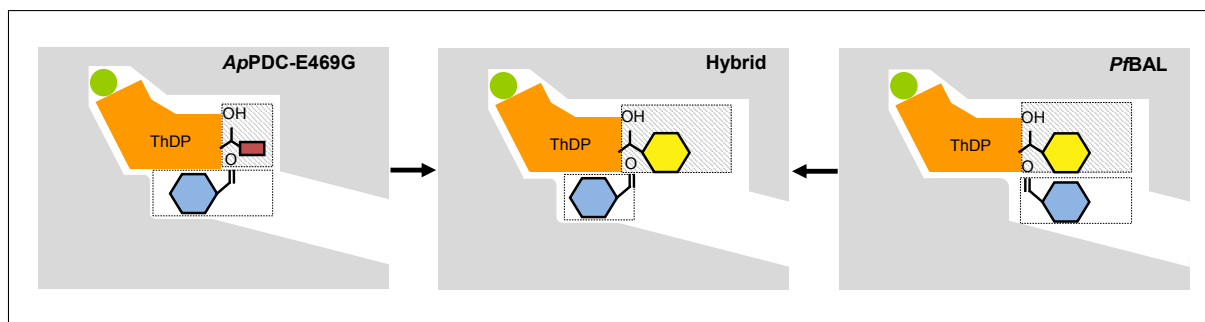


Figure 2.6: Schematic representation of the active site characteristics of *ApPDC*, *PfBAL* and a hybrid variant (derived from Westphal 2013). *ApPDC* variant E469G with its opened *S*-pocket provides a widened acceptor-binding site but a small donor-binding site. Carbonylation of acetaldehyde and benzaldehyde yields (*S*)-PAC (Rother et al. 2011), but formation of benzoin is hindered. In contrast, wt*PfBAL* provides a wide donor-binding site able to stabilize benzaldehyde. However, as acceptor, benzaldehyde is exclusively accepted parallelly oriented relative to the donor substrate due to a missing *S*-pocket, resulting in formation of (*R*)-benzoin. By combination of the enlarged acceptor-binding site of *ApPDC*-E469G and the wide donor-binding site of *PfBAL*, a hybrid could be constructed that facilitates binding of benzaldehyde in an antiparallel arrangement. Red rectangle: donor acetaldehyde, blue hexagon: acceptor benzaldehyde, yellow hexagon: donor benzaldehyde, green dot: Mg^{2+} ion, ruled box: donor binding site, white box: acceptor binding site.

In cooperation with Robert Westphal²⁷, a 'hybridization' approach was chosen as a strategy to provide access to (*S*)-benzoin (Figure 2.6). Two different hybrid enzymes were designed combining the large donor-binding site of *PfBAL* and the enlarged acceptor-binding site of *ApPDC*-E469G using either *PfBAL* or *ApPDC*-E469G as the template and changing the binding sites according to the respective other enzyme. Structural investigations of both enzymes, the *ApPDC* (pdb|2VBI, Rother et al. 2011) and the *PfBAL* (pdb|2AG0, Mosbacher, Müller, and Schulz 2005), combined with modeling of two benzaldehydes in antiparallel arrangement, revealed positions surrounding the active site cavities that prevent the required relative orientation of both benzaldehyde molecules for the formation of (*S*)-benzoin. Consequently, variants of *ApPDC*-E469G and *PfBAL* with widened donor- and acceptor-binding sites, respectively, were designed *in silico*. Subsequent expression and characterization of the variants was done by Robert Westphal.

Enlargement of the acceptor-binding site in *PfBAL*

Following a hybridization approach, engineering of *PfBAL* in order to design an enzyme for the direct synthesis of (*S*)-benzoin starting from benzaldehyde, means transplantation of the *S*-pocket of *ApPDC*-E469G into the active site of *PfBAL*. Comparison of both structures revealed

²⁷Robert Westphal, Institute of Bio- and Geosciences (IBG-1), Biotechnology, Forschungszentrum Jülich GmbH, Germany

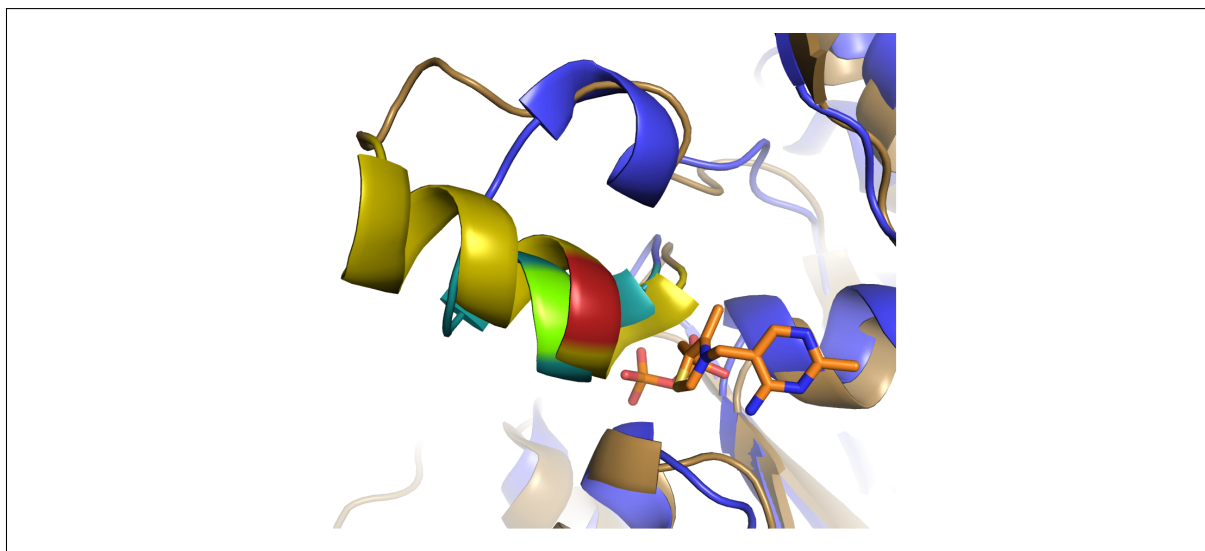


Figure 2.7: The structural superimposition of *ApPDC* (pdb|2VBI, blue) and *PfBAL* (pdb|2AG0, brown) by alignment of the ThDP cofactors revealed different lengths and minor deviations in the position of the α -helices PP- α E (cyan and gold). Due to those differences, *S*-pocket opening was not possible by a single mutation of T481 in *PfBAL* to glycine as compared to the E469G mutation in *ApPDC*. ThDP (orange) is exemplarily shown as resolved in the structure of *ApPDC*. Standard position 477, corresponding to E469 (green) in *ApPDC* and T481 (red) in *PfBAL*, is part of PP- α E.

a difference in the position of the protein backbone of an α -helix forming part of the active site pocket (Figure 2.7). The corresponding α -helix PP- α E (naming according to [Figure 4.5](#) on page 105) carrying E469 in *ApPDC* (477) is shifted towards the ThDP cofactor in the structure of *PfBAL* (pdb|2AG0, Mosbacher, Müller, and Schulz 2005), hindering introduction of a *S*-pocket by mutation of single positions. *In silico* mutation of T481G in *PfBAL* (477) increased the active site cavity (Figure 2.9 c/d on page 51) but as compared to the corresponding mutation in *ApPDC* (Figure 2.9 a/b), the effect was only marginal.

By mimicking α -helix PP- α E of the *ApPDC*-E469G variant in *PfBAL*, the helix should be shifted to a position as in the *ApPDC* structure, putatively opening a *S*-pocket. Thus, a variant was designed by adjusting PP- α E of *PfBAL* according to the respective helix of *ApPDC*-E469G (Figure 2.8 on the following page). By deletion of seven residues, the α -helix and the adjacent loop were shortened to mimic the corresponding region of *ApPDC*-E469G. Additionally, mutation T481G (477) putatively increased the distance between the protein surface and the ThDP cofactor, opening a *S*-pocket. Further substitution of the three residues 484-486 of *PfBAL* from 'FQQ' to 'IHD' according to the corresponding positions of *ApPDC* supported adaption of the structure of *ApPDC*-E469G. A structural model of the *PfBAL*/*ApPDC* hybrid sequence was generated using the homology modeling platform *SWISS-MODEL* (Arnold et al. 2006) with the crystal structure

of *PfBAL* (pdb|2AG0) as the template. As intended by the design, the model of the hybrid variant mimicked the active site cavity of *ApPDC*-E469G, therefore hypothetically opening the *S*-pocket (Figure 2.9 e on page 51). The distance of 5.5 Å between the C2 atom of the ThDP cofactor and the C α atom of the glycine residue at position 481 (477) of *PfBAL*-T481G was increased to a distance of 7.4 Å in the model of the hybrid variant.

<i>ApPDC</i> -E469G	457	IIFLINNRGYVIGIAIHD	-----	GPYNYIKNWDYAGLMEVF	492
Hybrid variant	469	IVIIMNNSWGAGLHIHD	-----	VTGTRLENGSYHGVAAAF	504
<i>PfBAL</i>	469	IVIIMNNSWGATLHFQQLAVGPNRVTGTRLENGSYHGVAAAF			511
		*:::***::: : :::		. . ::* .* *: .*	

Figure 2.8: Sequence alignment of the region containing α -helix PP- α E of *ApPDC*-E469G, *PfBAL* and the hybrid variant derived from a structural superimposition by STAMP (Russell and Barton 1992). Underlined residues are part of PP- α E. The hybrid variant was constructed using the sequence of *PfBAL*, deletion of 7 residues (-) and substitution of 4 residues (blue and yellow boxes) by the amino acids found at the respective positions in *ApPDC*-E469G. The blue highlighted position (standard position 477) corresponds to the position, which opened the *S*-pocket in *ApPDC* by the mutation E469G.


Cloning and expression of the gene resulted in soluble protein, which was successfully tested for carbonylation activity²⁸. However, the measured activity was weak (less than 1% conversion within 20 h) and the intended inversion of the enantioselectivity could not be observed. In contrast, (*R*)-2-HPP and (*R*)-benzoin were measured in enantiomeric excesses >99%. Subsequent MD simulation²⁹ in order to equilibrate the model of the hybrid variant provided the answer to the question, why the shift of PP- α E did not result in an inversion of the enzyme's enantioselectivity (Figure 2.9 f on page 51). The modified helix turned out to shift back closer to the cofactor as initially modeled by SWISS-MODEL, blocking the putative *S*-pocket. Nevertheless, the experiment revealed valuable information on the robustness of ThDP-dependent enzymes. Although only basal activity remained, deletion of seven and additional exchange of four amino acids being part of or being located in close vicinity to the active site did not completely eliminate

²⁸Cloning, expression and characterization of the *PfBAL* variant was done by Robert Westphal, Institute of Bio- and Geosciences (IBG-1), Biotechnology, Forschungszentrum Jülich GmbH, Germany. The gene of the novel hybrid variant was synthesized by GeneArt® Gene Synthesis (life technologies™, Carlsbad, USA), cloned into a pET19b vector (Novagen, Merck Millipore, Darmstadt, Germany) and overexpressed in *E. coli* BL21 (DE3) cells (Novagen, Merck Millipore, Darmstadt, Germany). Activity and enantioselectivity were analyzed in the homocoupling reaction of benzaldehyde and the cross-coupling reaction of benzaldehyde and acetaldehyde.

²⁹MD simulation of the hybrid variant were performed using GROMACS 4.5.4 (Spoel et al. 2005; Pronk et al. 2013). The variant was parametrized in the OPLS all-atom force field (Jorgensen, Maxwell, and Tirado-Rives 1996) and embedded in a cubic simulation cell with 3 Å distance to the protein surface. After filling the simulation cell with SPC/E water, 42 Na⁺ ions were added in order to neutralize the negative charge of the protein. Energy minimization was performed using the steepest decent algorithm. The variant was equilibrated for 1 ns and subsequently simulated for 5 ns.

its catalytic function.

Enlargement of the donor binding site in *ApPDC*

(For further details see publication 'A tailor-made chimeric thiamine diphosphate-dependent enzyme for the direct asymmetric synthesis of (*S*)-benzoins',  Section 4.3 on pages 119ff.)

A comparison of the donor binding sites of *ApPDC* (pdb|2VBI, Rother et al. 2011) and *PfBAL* (pdb|2AG0, Mosbacher, Müller, and Schulz 2005) with modeled antiparallel benzaldehydes revealed a threonine residue at position 384 in the *ApPDC* (388) to be the pivotal factor limiting the size of the donor-binding pocket. The analysis of the amino acid distribution in 186 sequences homologous to *ApPDC* revealed this threonine residue to be conserved in 92%, and 4% and 2% of the sequences having alanine or valine, respectively (Table 2.2 on page 52). In contrast, the respective position was found to be exclusively occupied by glycine in *PfBAL* and 42 homologs. Consequently, the variant *ApPDC*-E469G/T384G, combining the enlarged acceptor-binding site proven to allow for *S*-product formation (Rother et al. 2011) and the modified donor binding site mimicking the *PfBAL*, was tested for its potential to enantioselectively catalyze the homocoupling of benzaldehyde (for details see Section 4.3 on pages 119ff.)³⁰. With a conversion of 52% and a preference for (*S*)-benzoin (59% *ee*), the double mutant was the first known enzyme capable to (*S*)-selectively synthesize benzoin by direct asymmetric synthesis. Moreover, mutation T384G shifted the functional behavior of this enzyme from a PDC towards typical BAL activity. While acetoin and PAC were the only products of *ApPDC*-E469G in the carboligation of benzaldehyde and acetaldehyde, *ApPDC*-E469G/T384G catalyzed the reaction towards benzoin and 2-HPP.

In order to analyze the effect of mutant T384G on the structure of the variant, the structure was modeled by *in silico* mutagenesis using PyMol (*The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC, <http://pymol.org> 2013*) based on the structure of wt*ApPDC* (pdb|2VBI, Rother et al. 2011). Subsequent energy minimization and equilibration were applied using YASARA (YASARA Biosciences GmbH, Austria) to energetically optimize the structure of the variant. The standard procedure for energy minimization provided by YASARA was applied to a homodimer of the double variant including each one molecule of ThDP, a Mg²⁺ ion and two molecules of antiparallely arranged benzaldehyde per active site. A subsequent

³⁰Biochemical characterization was performed by Robert Westphal, Institute of Bio- and Geosciences (IBG-1), Biotechnology, Forschungszentrum Jülich GmbH, Germany

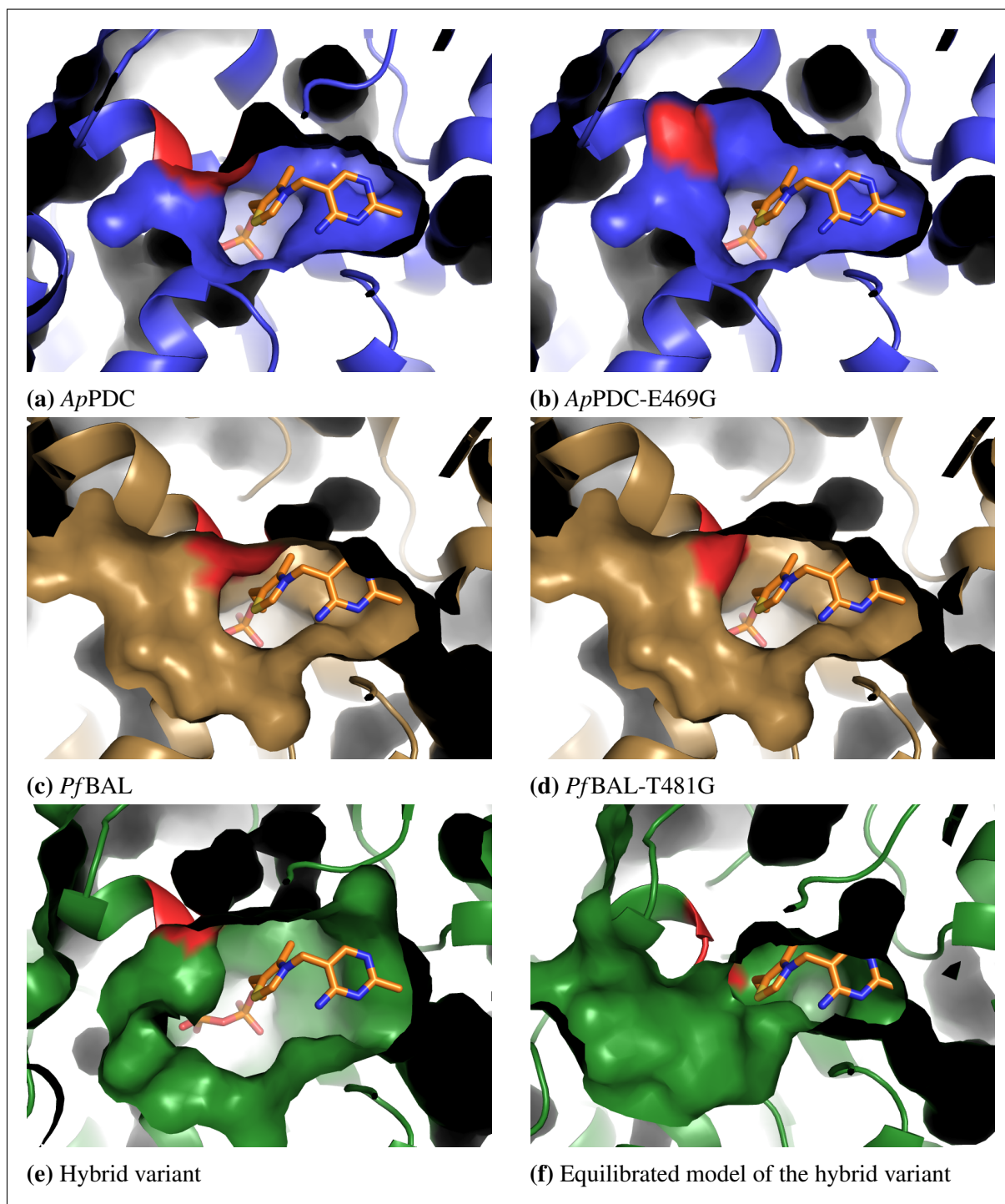


Figure 2.9: The superimposition of the structure of *ApPDC* (pdb|2VBI) (a), a model of *ApPDC-E469G* (b), the structure of *PfBAL* (pdb|2AG0) (c), a model of *PfBAL-T481G* (d), a model of the hybrid variant (e), and the presumed structure of the hybrid variant after equilibration (f) revealed differences in the position of the α -helix PP- α E (upper left corners). The deviating location relative to the ThDP cofactor hinders opening an *S*-pocket by mutation at standard position 477 (red) in *PfBAL*. Substitution of the glutamic acid residue at position 469 (477) in *ApPDC* opens the *S*-pocket (Rother et al. 2011) (b) as compared to the small cavity present in wt*ApPDC* (a). Mutation of the corresponding position in *PfBAL* increases the distance between the ThDP and the protein surface as well, but due to closer position of the protein backbone to the cofactor the resulting pocket is narrow (d). The model of the designed hybrid variant shows an increased cavity, but the equilibrated structure after MD simulation revealed a shift of the α -helix PP- α E towards the ThDP closing the *S*-pocket (f). Structures were superimposed according to the ThDP cofactor of *ApPDC* (orange).

simulation for 0.5 ns was appended in order to equilibrate the protein structure. For this purpose, the enzyme complex was embedded in a SPC/E water box with 2 Å distance to the protein surface and simulated using the YASARA2 force field (Krieger, Koraimann, and Vriend 2002). The equilibration revealed a tryptophane residue at position 388 (392) to rotate, additionally increasing the donor-binding pocket together with the space provided by the smaller side-chain of glycine as compared to threonine at position 384 (388).

The equilibrated structure of *Ap*PDC-E469G/T384G was further analyzed for positions interacting with and possibly stabilizing benzaldehyde bound in the original acceptor-binding site parallel to the donor substrate. Destabilization of substrate binding in the original acceptor binding site was previously shown to increase (*S*)-selectivity in the MenD enzyme from *E. coli* (Westphal, Hahn, et al. 2013). Visual inspection of the modeled structure of the *Ap*PDC-E469G/T384G

Table 2.2: Amino acid distribution at four enantioselectivity-determining positions in sequences homologous to *Ap*PDC and *Pf*BAL.

standard position	388	476	477	559
residue in <i>Ap</i> PDC ^[a]	T384	I468	E469	W543
residue in <i>Pf</i> BAL ^[b]	G393	A480	T481	L556
amino acid distribution in <i>Ap</i> PDC homologs ^[c]	92% T 4% A 2% V	95% I 2% T	97% E	76% W 4% F 4% L 4% S 3% R 3% T 2% E 2% G
amino acid distribution in <i>Pf</i> BAL homologs ^[c]	100% G	70% M 23% A 5% V	74% S 23% T	26% L 19% M 14% T 9% A 7% H 7% S 5% I 5% N 2% E 2% P 2% V

^[a] Position numbers are the absolute position numbers of the sequence of *Ap*PDC as contained in the crystal structure pdb|2VBI (Rother et al. 2011).

^[b] Position numbers are the absolute position numbers of the sequence of *Pf*BAL as contained in the crystal structure pdb|2AG0 (Mosbacher, Müller, and Schulz 2005).

^[c] Only amino acids with a minimum fraction of 1% are shown.

variant revealed an isoleucine residue at position 468 (476) and tryptophane 543 (559) to possibly stabilize parallel-oriented benzaldehyde through nonpolar interactions or π -stacking, respectively. Attenuation of such stabilizing effects potentially shifts the equilibrium to antiparallel-oriented benzaldehydes.

Sequence analysis showed the hydrophobic residues methionine, alanine and valine to occupy the corresponding position (standard position 476) in sequences homologous to *PfBAL* (Table 2.2 on the preceding page). Thus, in order to diminish stabilizing nonpolar interactions by increasing the distance between the parallel-oriented acceptor benzaldehyde and the side chain at position 468 (476) but to keep the conserved physicochemical property, isoleucine was replaced by valine, glycine or alanine. In fact, for all three variants increased (*S*)-selectivity (66-87% *ee*) was measured. The best variant *ApPDC*-E469G/T384G/I468A (87% *ee*, 95% conversion) was further optimized by site-saturation mutagenesis at position 543 (559) using NDT codon degeneracy, resulting in variant *ApPDC*-E469G/T384G/I468A/W543F with increased (*S*)-selectivity (95% *ee*) and moderate conversion (36%) in the formation of benzoin. Due to the variability in the amino acid distribution at standard position 559 (Table 2.2 on the preceding page), a prediction of a small set of promising substitutes was not possible. Therefore, usage of a random mutagenesis approach at this specific site was preferred over a more rational approach. For *meta*-substituted benzaldehydes, this variant reached excellent enantioselectivity (>99% *ee* (*S*)) with conversions ranging from 11% to 48%. Thus, detailed knowledge of the effects of different arrangements of donor and acceptor substrates on the enantioselective formation of chiral products, database analysis, modeling and equilibration of enzyme variants, and further optimization of an initial variant with already converted enantioselectivity allowed the rational engineering of a hybrid enzyme with the desired properties for enantiopure formation of various (*S*)-benzoin.

3 Discussion

3.1 Computing performance - a big issue in bioinformatics applications

In 1947, Margaret Dayhoff started to use punched-card machines to calculate molecular energies of organic molecules (Moody 2004; Strasser 2010). In order to point out her importance for the development of this scientific discipline, David Lipman, director of the NCBI since 1989, called her 'the mother and father of bioinformatics' (Moody 2004). Besides her role as a pioneer in the use of computers in chemistry, she further published the 'Atlas of Protein Sequence and Structure', the first database of proteins (Dayhoff et al. 1965), and developed software for the analysis of protein sequences. Therefore, she invented the famous Point Accepted Mutation (PAM) matrices, which were used to score sequence alignments of proteins.

With the development of fast sequencing methods revealing the nucleotide sequence of DNA, scientists were motivated to develop algorithms supporting the assembly of the obtained sequence fragments into genes and complete genomes. Frederick Sanger, who lent his name to the Sanger sequencing technique, used an IBM mainframe to execute a program assembling so-called sequence shotguns into larger sequences (Moody 2004). Rodger Staden, who worked in the group of Sanger, took a key step in the development of bioinformatics when he brought shotgun assembly to computers directly in the lab in 1979. To abolish the spatial separation of sequencing and sequence assembly, he wrote a program to be executed on PDP-11 computers available in the lab (Moody 2004). From that point on, computers in laboratories were more and more applied to analyze data in genomics and proteomics, which did accelerate development of specific algorithms. In 1985 and 1988, Lipman and Pearson published their tools FASTP and FASTA for comparisons of protein and DNA sequences (Lipman and Pearson 1985; Pearson and Lipman 1988). To improve performance by use of heuristics and to include statistics about

the significance of results generated by an algorithm aiming for sequence comparison, Lipman and coworkers subsequently also presented the BLAST program in 1990 (Altschul et al. 1990). Benefiting from the World Wide Web, which became available at that time, BLAST became the most prominent sequence search tool in bioinformatics (Moody 2004).

Besides development of improved algorithms, bioinformatics profited from the improvements in the performance of the executing computers. One of the factors determining the performance of a central processing unit (CPU) is its clock rate¹. The clock rate describes the frequency at which a CPU is executing instructions. As exemplarily shown for CPUs manufactured by Intel², the clock rate was rapidly increased from 740 kHz (Intel 4004) to 4 GHz (Intel i7-3970X) between 1978 and 2011 (Figure 3.1), resulting in an increase in the CPU performance. In the course of this development, the architectures were changed from CPUs with one core to multi-core processors in order to reduce energy consumption and waste heat emission. Thus, in order to achieve the calculation performance available with modern CPUs, the manufactures additionally implemented the performance-gaining parallelization support within single chips.

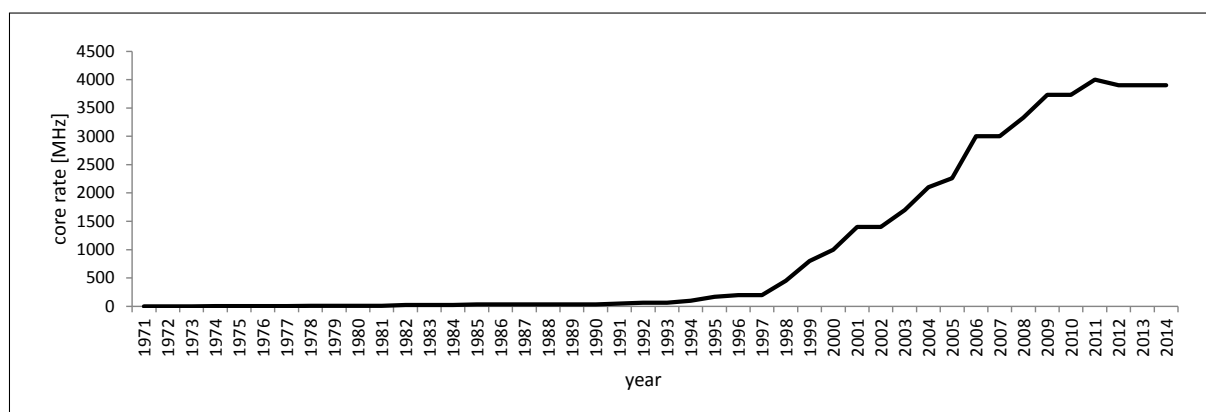


Figure 3.1: Development of the clock rate of Intel CPUs as listed at Wikipedia (http://en.wikipedia.org/wiki/List_of_Intel_microprocessors). The clock rates of Intel CPUs were continuously increased since release of the first processors in 1971. In the years between 1996 and 2011, the frequency was increased by a factor of 20 from 200 MHz to 4 GHz. However, in the years since 2009, the progression in the development of CPUs with higher clock rates decelerated.

Although the performance of available CPUs was continuously improved during the last decades, the increase in the clock rates is not able to keep up with the drastic increase in the number of available protein sequences. Within ten years (2000-2010) the core frequency of CPUs manufactured by Intel was quadrupled whereas the number of known protein sequences in the

¹Further factors influencing the performance are the memory clock rate, the front-side bus speed, availability and size of CPU cache, and others.

²Intel Corporation, 2200 Mission College Blvd., Santa Clara, CA 95054-1549, USA

UniProt/TrEMBL (The UniProt Consortium 2014) increased by a factor of 160 in the same period (Figure 1.2 on page 4). As a consequence, the soft- and hardware developed to manage the amount of available sequence data in the first years of the 21st century cannot further be applied nowadays. As an alternative, performance-optimized versions of previously developed programs or programs with newly designed algorithms emerged. *Blastp*, an advancement of the initial BLAST program for the search for homologous protein sequences, was accelerated by optimizing usage of the L2 cache³ and distribution of calculations on multiple available CPUs and/or CPU cores (Camacho et al. 2009). By the use of more heuristic algorithms and in combination with parallelization, *usearch* additionally reduced the time needed to search for homologous sequences as compared to BLAST by a factor of 350 (Edgar 2010). In the same way programs for multisequence alignments (e.g. *clustalo*, Sievers et al. 2011), analysis of DNA sequences (e.g. *U-BRAIN*, D'Angelo and Rampone 2014), analysis of RNA sequences (e.g. *sPARTA*, Kakrana et al. 2014) and molecular dynamics simulation applications (e.g. *GROMACS*, Pronk et al. 2013) were optimized for performance.

The TEED, which was used as the basis of many analyses done in this work, was initially created by Michael Widmann and co-workers in 2009 (Widmann, Radloff, and Pleiss 2010) using the *DWARF* system developed in the years before 2004 (Fischer 2004). Due to the enormous increase in the number of available sequences, the *DWARF* system and the previously available computational hardware were no more suited to update this FSPD to be able to serve as the basis for up-to-date research. Thus, the development of the scalable *DBParse* and *DBUpdate* algorithms, both extensively exploiting the potential of bioinformatical sequence analysis to be parallelized and implementing performance-optimized versions of search and clustering tools like *usearch* (Edgar 2010) and *blastp* (Camacho et al. 2009), was pivotal for this work. By concomitant optimization of the underlying relational data model and the simultaneous development of the *BioCatNet* system including a new, optimized graphical user interface, a successor of the *DWARF* system was established. The *BioCatNet* system proved its applicability and houses FSPDs for ThDP-dependent enzymes (TEED) (Vogel and Pleiss 2014), imine-reductases (IRED) (Scheller et al. 2014), Cytochrome P450 monooxygenases (CYPED)⁴, transaldolases (TALED)⁴, multicopper oxidases (LacED)⁴, laccases (LccED)⁴ and others⁴.

³Modern CPUs possess small but fast memory to cache data repeatedly requested from the system's main memory to accelerate access to this data. The cache is hierarchically organized in levels with increasing speed and latency (L1, L2, ...).

⁴The mentioned versions of these databases implemented in the *BioCatNet* system were not published so far.

Besides parallelization and algorithm optimization, bioinformatics follows further strategies to improve computing performance including the usage of application-specific integrated circuits (ASIC) (Shaw et al. 2008) or graphics processing units (GPU) (Hallock et al. 2014) instead of the more generalized CPUs. In addition, computing performance can be increased by optimization of the speed of read and write events to or from an applied storage medium. In this work, virtual storage devices (referred to as 'RAM disks') were mounted in the applied servers in order to execute processes with frequent read and write accesses in the volatile system's main memory (random-access memory (RAM)). Latencies due to writing and reading to/from temporary files during generation of databases and systematic, large-scale analysis of sequence data were thereby reduced to a minimum.

3.2 The sequence-structure-function relationships of ThDP-dependent enzymes

3.2.1 Sequence-Structure

Standard numbering schemes

The sequence-structure relationship can be summarized as the trait of the primary structure to determine the three-dimensional structure, in defiance of the indeterminacy of most positions concerning the occurrence of defined amino acids. As repeatedly mentioned, the structure of enzymes is more conserved within enzyme families as compared to the sequences. This is founded in the functional role of enzyme's structures that assemble individual functionally relevant amino acids and ligands in specific arrangements that enable the various catalytic effects. In this regard, the exact position of specific amino acids on the primary structure is of less importance. However, the backbone atoms of amino acids at corresponding positions in the sequences of evolutionary related enzymes are commonly located at the same positions in the structures (in this thesis referred to as 'structurally equivalent residues'). The standard numbering schemes developed as part of this work rely on this relationship between the sequence and the structure. Prediction of structural equivalence becomes feasible just by sequence alignments and the use of structure information from a few representative members of an enzyme family. Moreover, it enables comparison of results obtained with variants of different enzymes but at

equivalent positions (Table A.3 on pages 177ff.). Thus, this method is highly beneficial for an analysis and reinterpretation of published (and in the best case even via *BioCatNet* electronically accessible) results on variants and thus for planning of new variants aiming for the design of enzymes with desired properties. By taking available information on the functional roles of certain positions of a protein family into account, the method allows the prediction of functionally or structurally relevant residues as discussed later.

Although originally developed for the **ThDP-dependent DCs** but designed to be applicable to other protein families, the standard numbering method was transferred to the **entire family of ThDP-dependent enzymes** as well as to Cytochrome P450 monooxygenases, laccases and triterpene cyclases. Prerequisite for a reliable domain-guided alignment is a structural alignment of a sufficiently high number of representative structures in order to cover the family-specific sequence space. However, as shown for the domain-based standard numbering scheme for the entire family of ThDP-dependent enzymes, manual extension of an initial structural alignment by addition of further sequences allows to widen the specificity of the HMM-profiles to subfamilies without available structure information. Thus, the method provides the required flexibility for a broad applicability to further protein families.

In cooperation with the curator of the CYPED, Łukasz Gricman⁵, standard numbering schemes for the highly divergent protein family of **Cytochrome P450 monooxygenases (CYP)** were generated (For further details see publication 'Gricman, Vogel, and Pleiss 2014. Conservation analysis of class-specific positions in Cytochrome P450 monooxygenases: functional and structural relevance. *Proteins* **82**:491-504, doi:[10.1002/prot.24415](https://doi.org/10.1002/prot.24415)' (not attached to this thesis)). Based on information about the family organization and representative structures provided by the curator, structural alignments were generated and HMM profiles were derived according to the strategy described for the ThDP-dependent DCs⁶. With the most recent update of the CYPED, Łukasz Gricman assigned the different proteins to two classes (I and II) based on literature information (Gricman, Vogel, and Pleiss 2014). Due to higher sequence and structure similarity within the classes I and II as between the two different classes and in order to facilitate accurate standard number assignment, two independent class-specific standard numbering schemes were developed. Similar to the validation done for the standard numbering scheme for ThDP-dependent DCs, a comparison of alignments based on the CYP standard numbering

⁵Institute of Technical Biochemistry, University of Stuttgart, Germany

⁶For details on the method see Section 2.2.1 on page 32

schemes using *nvwAlign* and *STAMP* was done. The comparison of the class I and class II alignments obtained from the different alignment methods revealed 86% and 81% of the columns, respectively, to be identical.

After the regeneration of the **Laccase and Multicopper Oxidase** Engineering Database (LccED) by Silvia Fademrecht⁷ using the *DBParse* tool (see Section 2.1.2 on page 23), a standard numbering scheme was generated for this protein family. Similar to the ThDP-dependent enzymes, multi copper oxidases (MCO) from different superfamilies deviate in their composition of domains. An in-depth analysis of the modular structure of laccases done by Silvia Fademrecht revealed three domains (N: N-terminal domain, M: middle domain, C: C-terminal domain) in N-M-C order to constitute the sequences of a majority of MCOs. However, subfamilies were identified exclusively consisting of the domains N and C or more complex arrangements of three subsequent N domains connected by an additional domain followed by a terminating C-domain. Consequently, a domain-based standard numbering scheme was developed for the family of MCOs and laccases following the procedure described for the domain-based standard numbering scheme for ThDP-dependent enzymes (see Section 2.2.2 on pages 37ff.). Using a selection of representative structures from seven superfamilies, structure-based alignments for the three individual domains N, M and C were created and used to derive domain-specific HMM profiles. Using the HMM profiles, standard numbers were assigned to the respective domains of all MCOs using the sequence of the *Trametes versicolor* laccase as the reference. Subsequently, the numbered domains and the derived multisequence alignments⁸ served Silvia Fademrecht as the basis for conservation analyses and analyses of the inner-domain substrate binding loops of MCOs⁹.

Further, a numbering scheme was developed for the family of **triterpene cyclases** in cooperation with Silvia Fademrecht¹⁰ using the same methodology. The sequence of the squalene-hopene cyclase from *Alicyclobacillus acidocaldarius* (*AacSHC*) was defined to be the reference sequence since it is the best-documented representative of the triterpene cyclase family. Beside the *AacSHC* (pdb|1UMP, Reinert, Balliano, and Schulz 2004), structure information is also available for the human oxidosqualene cyclase (*HsaOSC*, pdb|1W6K, Thoma et al. 2004). Both structures were

⁷Silvia Fademrecht, Institute of Technical Biochemistry, University of Stuttgart, Germany; The updated version of the TTCED (Racolta et al. 2012) is not published yet.

⁸Multisequence alignments were generated using the *nvwAlign* program developed as part of this work.

⁹Fademrecht, S; Vogel, C; Le Roes-Hill, M; Pleiss, J. A systematic analysis of subfamily-specific properties of Multicopper Oxidases. *Manuscript in preparation*.

¹⁰Silvia Fademrecht, Institute of Technical Biochemistry, University of Stuttgart, Germany

superimposed by *STAMP* and the derived alignment was expanded by 47 sequences using sequence-to-profile alignments by *clustalw2* (Larkin et al. 2007). The 47 additional sequences were selected using *usearch* (Edgar 2010) by clustering the available sequences of triterpene cyclases with an identity threshold of 30%. The longest sequence of each cluster was taken as representative. The subsequently deduced HMM profile was used to assign standard numbers to all sequences of an updated version of the Triterpene Cyclase Engineering Database (TTCED)¹¹. Systematic analysis of the families of SHCs and OSCs using the assigned standard numbers is ongoing in the project of Silvia Fademrecht.

Modularity of ThDP-dependent enzymes

ThDP-dependent enzymes from different superfamilies deviate in their family-specific sequence architecture (Figure 2.2 on page 38). Consequently, enzymes from different superfamilies deviate enormously according to their global sequence similarity. However, even when compared locally by exclusively measuring the sequence similarity of the catalytically most relevant PYR and PP domains, differentiation between members of different superfamilies was observed (Figure 4.6 on page 108). Moreover, ThDP-dependent enzymes from different superfamilies deviate considerably concerning their overall structures. Due to combination of the PYR and PP domains with additional domains and different linkage of the individual domains, the available enzyme structures from different superfamilies can hardly be superimposed automatically. In detail, even the PYR and PP domains show family-specific differences (see [Figure 4.2](#) Section 4.2.4 on pages 104ff.). Thus, ThDP-dependent enzymes were evolutionary separated based on their catalytic function and their structural architecture ([Figure 4.8](#) on page 115).

However, manual alignment of the available structures of this enzyme family revealed a conserved structure of the common catalytic core. By classifying the different structures based on the order of the PYR and PP domains on the sequence, if both domains were coupled at all, and the location of the ThDP cofactor between PYR and PP domains from the same or different monomers, only five different structural architectures were observed ([Figure 4.3](#) on page 100). Moreover, independent from the respective architecture, the location and relative orientation of the PYR and PP domains forming the active sites resemble in all available enzyme structures. Although the global sequences and structures of ThDP-dependent enzymes vary considerably,

¹¹The TTCED is curated by Silvia Fademrecht, who updated the database using the *DBParse* tool described in this work (see Section 2.1.2 on page 23).

the sequences of the PYR and PP domains determine a well-defined structure of both domains and of the common catalytic core. To conclude, the architecture of the domains on the primary structure as well as the acquisition of further domains only have marginal effects on the core structure.

In order to prove this concept, the sequence of the *ApPDC* was modified (for details see Section 2.3.2 on pages 41ff.) to convert the enzyme from an inter-monomer/PYR-PP architecture into an intra-monomer/PP-PYR type enzyme. In order to investigate the effects of this redesign on the folding into a soluble protein, the catalytic activity and the actual structure of the active complex, an experimental characterization was done by Saskia Bock¹². The obtained preliminary results hardly allow to draw conclusions, since a SDS gel electrophoresis comparing the crude extracts of cells transfected with an empty vector and a vector carrying the design gene, respectively, showed overexpressed protein with a mass of approximately 40 to 43 kDa for both samples after 24h (Figure 2.4 on page 43). Since the successful cloning was approved by sequencing of the vector, the results suggest a failing protein expression or degradation of putatively misfolded protein. Thus, the correct folding of the designed construct into a soluble protein has not been shown so far. Consequently, the assumed potential of the PYR and PP domains of the *ApPDC* to fold into the ThDP-binding fold (Figure 4.5 on page 105) independent from their order on the sequence and to concertedly form active sites, has not been approved yet. However, optimization of the variant by mutation of the two flexible regions on the protein surface detected by MD simulation (Figure 2.3 on page 42) and co-expression of chaperons might enable expression of soluble protein. Thus, this project is ongoing as part of the doctorate of Saskia Bock.

3.2.2 Structure-Function

Structurally equivalent residues

In this thesis, the term 'structurally equivalent residues' is used for amino acids from different proteins, having their backbone atoms superimposed in a structure alignment. This definition is arbitrary, since 'structural equivalence' does not necessarily have to be defined by the position of the amino acid backbone. It could contrarily describe amino acid side chains from different proteins and with deviating backbone positions that occupy the same space in the three-dimensional

¹²Institute of Bio- and Geosciences (IBG-1), Biotechnology, Forschungszentrum Jülich GmbH, Germany

structure of the respective protein (Figure 3.2).

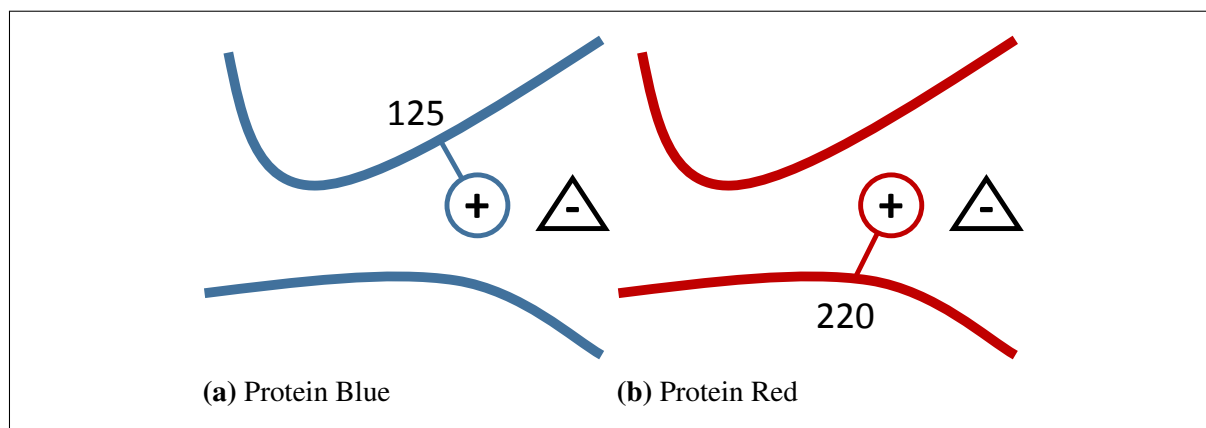


Figure 3.2: Structural equivalence of functionally relevant amino acid side chains with different standard position numbers in two different proteins. a) A residue with positively charged side chain at standard position 125 serves as an anchor for a negatively charged ligand. b) In the homologous protein 'Red', the residue at standard position 220 with identical physicochemical properties as residue Blue-125 adopts the role as the anchor.

With the family-specific standard numbering schemes presented in this thesis, addressing and analysis of functionally and structurally relevant residues is possible. However, the method is limited to positions that are equivalent in their $C\alpha$ position. Thus, the method misses such positions that are located differently concerning their backbone atoms although sharing their functional role by having a common effect in the same three-dimensional space in the protein structure. Functionality of enzymes is conferred by an enzyme-specific spatial arrangement of amino acids, cofactor and water molecules as well as metal ions. Moreover, the exact orientation of the amino acid backbone and side chains plays a role for the activity of enzymes and the structural integrity.

In ThDP-dependent enzymes, several positions were described to play a structural or functional role via their backbone carbonyls (*Geobacillus stearothermophilus* PDH E1-I206, Frank, Titman, et al. 2004; Frank, Pratap, et al. 2005; *Gs*PDH E1-E126, Pei et al. 2008; *Zm*PDC-G469, Candy and Duggleby 1998; *Zm*PDC-G413, Lie et al. 2005; Kluger and Tittmann 2008; *Sc*PDC-E148, Jordan, Nemeria, et al. 1998; *Lp*POX-A420, Meyer, Neumann, Koers, et al. 2012) and backbone amides (*Thermus thermophilus* SPDC-N154/S155, Graham et al. 2009; *Lp*POX-G35/S36, Kluger and Tittmann 2008; *Zm*PDC-D27, Meyer, Neumann, Parthier, et al. 2010). Beyond that, the majority of positions playing a role in substrate activation, chemo- and enantioselectivity of ThDP-dependent enzymes deploy an effect by their functional side chains. Consequently, although the amino acid side chains theoretically can obtain different structural orientations (referred to as rotamers), specific rotamers are preferred at specific positions. Also the orientation

of active carbonyls or amides of the peptide backbone is well-defined for such positions playing a functional or structural role. Thus, having structure information of enzymes available, analysis of the spacial equivalence of physicochemical similar functional moieties in different homologous proteins would provide a more complete view on functionally relevant residues.

Within the scope of her bachelor thesis, Chantal Göttler developed a tool¹³ to discover the occurrence of hydrogen bond donors and acceptors of different proteins affecting the same space in the three-dimensional structures (Göttler 2013). Using a simplified model of interactions via common 'effect points', the method was designed to predict positions of different proteins with different backbone positions but putatively shared functional role. Due to the strikingly high number of resulting hits of such pairs of positions and lack of a scoring function that could be used to rank the hits by their potential functional relevance, this method was not productively applied in this thesis. Moreover, parametrization of hydrophobic and ionic interactions as well as disulfide bonds is missing. After completion of the parametrization and integration of an adequate scoring function, this method might in future support the identification of functionally or structurally relevant residues based on available structure information.

Functional immunity of ThDP-dependent enzymes against mutations

As shown for various representatives of this protein family, mutations in the active site mostly are non-fatal for the activity. Even a drastic variation of helix PP- α E contributing to the active site of *PfBAL* (see Section 2.4.1 on page 47) did not completely extinguish catalytic activity. Furthermore, more directed mutagenesis intended to modulate chemo- and enantioselectivity of *ApPDC* (Rother et al. 2011; Westphal et al. 2014a; Westphal 2013), *EcMenD* (Westphal, Hahn, et al. 2013; Westphal, Waltzer, et al. 2013), the MenD from *Bacillus subtilis* (*BsMenD*); Dawson et al. 2010; Westphal, Jansen, et al. 2014), the acetohydroxyacid synthases I and II from *E. coli* (*EcAHAS I/II*; Engel et al. 2003; Tittmann, Vyazmensky, et al. 2005; Hill and Duggleby 1998; Steinmetz et al. 2010; Vyazmensky et al. 2011; Belenky et al. 2012; Schmitz 2012), *ZmPDC* (Bruhn et al. 1995; Schenk, Leeper, et al. 1997; Candy and Duggleby 1998; Iding, Siegert, and Pohl 1998; Pohl, Siegert, et al. 1998; Chang, Nixon, and Duggleby 1999; Wu et al. 2000; Huang et al. 2001; Lie et al. 2005; Siegert et al. 2005; Wechsler 2014; Meyer, Walter, et al.

¹³*SERgrid* - a tool for the analysis of structurally equivalent residues. For details see Göttler 2013 and Supporting Informations, Section A.2 on page 175.

2011), *EcTK* (Galman et al. 2010) and many others¹⁴ proved the durability of those enzymes against 'open heart surgeries'. To exaggeratedly simplify: as long as the cofactor is kept intact, the enzyme retains activity. However, mutations in the active sites of the previously mentioned enzymes caused significantly reduced activity in most instances. This shows clearly, which role the cofactor and the surrounding protein have. Besides optimal binding and arrangement of ThDP in the 'V-conformation', as well as activation and recycling of the cofactor, the protein plays no further relevant role for the activity of the enzyme complex¹⁵. The enzyme's duty is to control substrate specificity, regulation of the activity by activation or inhibition, as well as the enantioselective formation of chiral products. As shown in the literature and by the variants designed as part of this thesis, regulation of chemo- and enantioselectivity is in many ThDP-dependent enzymes simply solved by steric limitations. By providing substrate-binding pockets sterically optimized for specific substrates and specific arrangements of the donor and acceptor in ligation reactions, different enzymes possess defined substrate- and enantioselectivity. This makes ThDP-dependent enzymes convenient targets for rational engineering and enabled the generation of an enzyme toolbox for the synthesis of α -hydroxy ketones. Consequently, the majority of mutations in ThDP-dependent enzymes were done within or in proximity of the active sites (Figure 3.3 on the following page). This indicates that the shape of the substrate binding pockets, and thus the structure of ThDP-dependent enzymes, highly influences the function (structure-function relationship).

3.2.3 Sequence-Function

Standard numbering schemes

As already mentioned, the standard numbering schemes presented in this thesis are able to predict the location of the backbone atoms of residues in the three-dimensional structure. This ability arises from the exploitation of the sequence-structure relationship. By making use of the structure-function relationship, the method can further be applied to identify functionally relevant residues without need for structure information about the respective enzyme. For positions with common backbone positions and conserved amino acids or at least conserved physicochemical

¹⁴For a comprehensive list of variants of ThDP-dependent enzymes see Supplementary Information, Section A.3 on pages 177ff.

¹⁵Recent studies on the conformation of ThDP in different ThDP-dependent enzymes further indicate a role of the enzyme in facilitating out-of-plane distortion in the aminopyrimidine and the thiazolium moieties, which triggers enzyme-specific mechanistic behavior (Meyer, Neumann, Koers, et al. 2012).

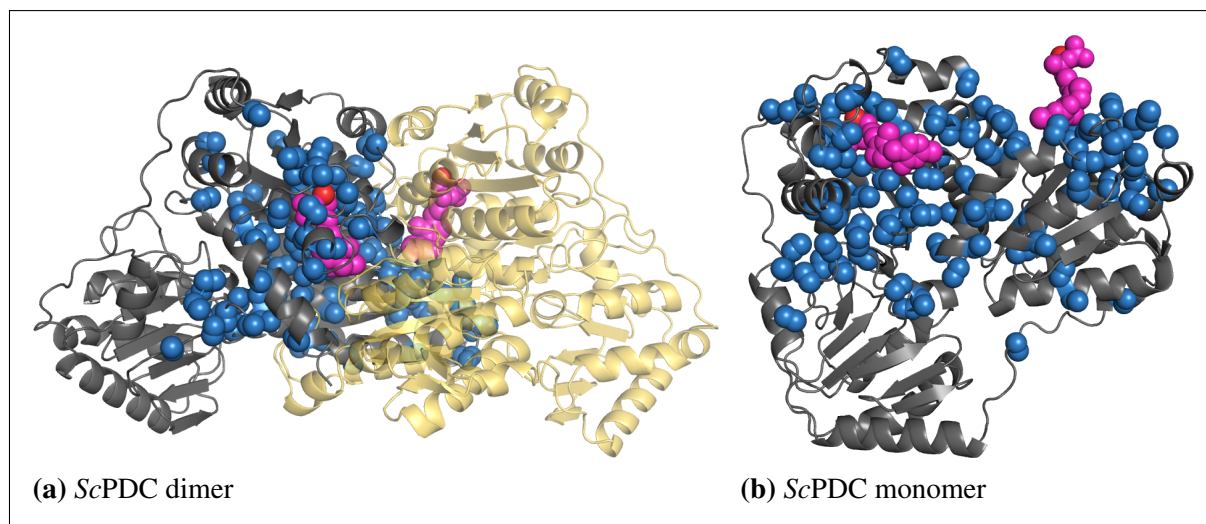


Figure 3.3: Representation of the location of positions, that were mutated in ThDP-dependent enzymes from different superfamilies and were documented in the literature (For a list of those positions see table A.3 on pages 177ff.). The positions were mapped onto the structure of *ScPDC* (pdb|2VK8, Kutter et al. 2009) using the standard numbering scheme for ThDP-dependent enzymes. (a) The two monomers forming the active dimer of *ScPDC* (black, yellow) as well as the Mg^{2+} ions and ThDP molecules (red and pink) bound in the two active sites are colored differently. The $C\alpha$ and $C\beta$ (if available) atoms of residues, which were found in literature to have been varied, are shown as spheres (blue). Notably, the majority of the mutated positions is located in close proximity to the two active sites.

properties, common functional roles stand to reason. Thus, the standard numbering schemes allow prediction of functional roles of single amino acids based on a limited set of structurally determined family members, an adequately accurate sequence alignment method, the pure amino acid sequence of the enzymes of interest, and information on the functional roles of certain positions in representatives from the same family.

Exploiting the knowledge about the sequence-function relationship of ThDP-dependent enzymes for rational engineering

In proteins sharing highly similar sequences and structures, the sequence-function relationship naturally resembles the structure-function relationship. However, the ThDP-dependent enzymes from different superfamilies deviate enormously in respect to their global sequences and structures. Nevertheless, the standard numbering scheme for ThDP-dependent enzymes clearly demonstrates the functional resemblance of residues with the same standard numbers. The capability to assign common standard numbers to structurally equivalent and thus putatively functionally similar positions results from the strikingly well conserved structure and relative orientation of the catalytically relevant PYR and PP domains and the existence of highly con-

served residues on the sequences of both domains, which serve as anchors in the profile-guided sequence alignments.

The high accuracy in the prediction of structurally equivalent residues using only the sequences of ThDP-dependent enzymes as an input allows to apply the method for the prediction of positions with putative functional relevance. Using information from literature about positions lining the *S*-pocket¹⁶ (Gocke 2007) and the standard numbering scheme, the *S*-pocket forming residues of *EcMenD*, *BsMenD*, *ApPDC* and *PfBAL* were predicted (Table 3.1 on pages 68ff.). Since the predicted positions line the active site cavities of those proteins and were shown to often have drastic effects on the function of the concerned enzymes, those positions had a good chance to be hotspots for the modulation of the enantioselectivity. Especially, since mutations at standard positions 28, 476, 477 and 480 were reported to have the potential to invert enantioselectivity for certain reactions (Table 3.1 on page 68). By molecular modeling in collaboration with Robert Westphal, the effects of different positions in *EcMenD*, *BsMenD*, *ApPDC* and *PfBAL* were evaluated, promising variants were designed and subsequently experimentally characterized¹⁷.

In order to invert the enantioselectivity of the strictly (*R*)-selective (>99% *ee*) *EcMenD* and *BsMenD*, Robert Westphal and co-workers mutated both enzymes at standard positions 476 and 477 (Westphal, Waltzer, et al. 2013; Westphal, Hahn, et al. 2013; Westphal, Jansen, et al. 2014). The variants I474A(476)/F475G(477) of *EcMenD* and I489A(476)/F490A(477) of *BsMenD* both showed inverted enantioselectivity (75% *ee* (*S*) and 92% *ee* (*S*), respectively) in the carboligation of α -ketoglutarate and benzaldehyde¹⁸. In accordance with the *S*-pocket concept, these mutations opened up space needed for benzaldehyde to bind in antiparallel orientation relative to the donor substrate α -ketoglutarate.

Due to a shift of α -helix PP- α E in *PfBAL* as compared to other ThDP-dependent DCs, single mutations of the predicted *S*-pocket residues are not sufficient to open up an *S*-pocket (see Section 2.4.1 on page 47). Moreover, comprehensive redesign of the respective region did also not lead to an inverted enantioselectivity. Thus, (*S*)-selective carboligation using variants of *PfBAL* as biocatalysts remains an unsolved challenge. However, by mutation of the *S*-pocket lining residues of the *ApPDC* (I468A and E469G at standard positions 476 and 477, respectively)

¹⁶For a comprehensive list of positions mutated in ThDP-dependent enzymes see Supporting Information, Section A.3 on pages 177ff.

¹⁷Experimental characterization of all enzyme variants was done by Robert Westphal, Institute of Bio- and Geosciences (IBG-1), Biotechnology, Forschungszentrum Jülich GmbH, Germany

¹⁸Further variants were generated and tested but the respectively highest (*S*)-selectivities were obtained using *EcMenD*-I474A(476)/F475G(477) and *BsMenD*-I489A(476)/F490A(477).

complemented by enlargement of the donor-binding site (T384G, standard position 388), a variant capable to catalyze the direct asymmetric synthesis of (*S*)-benzoin starting from benzaldehyde was tailored. As shown for the *S*-pocket lining residues, standard position 388 is likewise well documented in the scientific literature due to its effect on the binding of different substrates (see Supporting Information, Table A.1 on page 177). In the indole-3-pyruvate decarboxylase from *Enterobacter cloacae*, the PDC mimicking substitution of glutamine 383 (388) by threonine resulted in an increased activity for pyruvate and the physiological substrate 3-indolepyruvate but decreased the activity for benzoylformate conversion (Schütz, Golbik, et al. 2005). Mutations at the corresponding positions of the catalytic subunits of *EcAHAS-I* and *EcAHAS-II* likewise showed an influence on the substrate specificity (Belenky et al. 2012; Steinmetz et al. 2010). In order to allow binding of pyruvate as the donor substrate, *EcGXC* variant I393A was generated, which successfully converted the glyoxylate carboligase into an acetolactate synthase (Nemeria, Binshtein, et al. 2012). Thus, the examples from literature and the results obtained in this thesis document that positions at the same standard position on sequences of different ThDP-dependent enzymes have the same functional impact. Based on that, standard numbering schemes make prediction of hotspots with functional impact feasible from the sequence itself without need for detailed biochemical characterization or structure determination.

Table 3.1: Selection of positions predicted to be part of the *S*-pockets of the respective proteins and variants of different ThDP-dependent DCs at those positions^a. For a more comprehensive list of variants collected from literature see Supporting Information, Table A.1 on page 177.

Std. pos. ^b	<i>EcMenD</i>	<i>BsMenD</i>	<i>ApPDC</i>	<i>PfBAL</i>	Variants ^c	Effect	References ^d
28	S32	S31	D27	A28	<i>ECIPDC</i> - D29E	decreased activity, pyruvate and benzoylformate are still accepted as substrates, but not indolepyruvate	[1]
					<i>PfBAL</i> - A28S	decreased activity, accepts benzoylformate for decarboxylation which wt does not	[2,3,4,5]
					<i>PpBFDC</i> - S26A	increased k_m and decreased k_{cat} in decarboxylation of benzoylformate	[1,3,6]
					<i>ScPDC</i> - D28A/N	position is involved in allosteric activation, variants produce acetolactate; D28A produces (<i>R</i>)-PAC	[7,8,9,10,11,12]
					<i>SvPPDC</i> - S25D/N	increased k_m and decreased activity in decarboxylation of phosphonopyruvate	[13]
					<i>ZmPDC</i> - D27A	affinity for ThDP decreased, variant produces acetolactate	[14]
393	R395	R409	F389	L398	<i>ZmPDC</i> - D27E/N	activity in production of acetaldehyde decreased	[15]
					<i>BsMenD</i> - R409A	involved in binding of α -ketoglutarate and isochorismate, increased k_m , decreased activity	[16]
476	I474	I489	I468	A480	<i>NtAHASI</i> - N489V	decreased affinity for FAD, inactive variant	[17]
					<i>KdcA</i> - V461I	increased k_m and decreased activity for 3-methyl-2-oxopentanoic acid and phenylpyruvic acid but decreased k_m and increased activity for benzoylformic acid and pyruvate	[18]
					<i>BsMenD</i> - I489A/G	involved in binding of α -ketoglutarate and isochorismate, increased k_m , decreased activity	[16]

continued Table 3.1: Variants of ThDP-dependent DCs at positions lining the S-pocket.

Std. pos. ^b	<i>EcMenD</i>	<i>BsMenD</i>	<i>ApPDC</i>	<i>PfBAL</i>	Variants ^c	Effect	References ^d
					<i>EcAHASI</i> -L476M	increased k_m for pyruvate and decreased activity in formation of acetolactate	[19]
					<i>EcGXC</i> -L478A	decreased activity in tartronate semialdehyde formation	[20]
					<i>PfBAL</i> -A480I	increased k_m for 2-HPP, decreased activity in cleavage of 2-HPP and benzoin	[5]
					<i>PpBFDC</i> -A460G	increased k_m and decreased activity for benzoylformate, little effect on enantioselectivity	[21,22]
					<i>PpBFDC</i> -A460Y/I	decreased activity for benzoylformate and increased activity for aliphatic substrates; enantioselectivity was inverted to (<i>R</i>)-2-HPP in double mutant A460I/F464I	[22,23,24]
					<i>ZmPDC</i> -I472A/F/L/S	decreased activity in decarboxylation of pyruvate and PAC formation; and reduced (<i>R</i>)-selectivity in PAC formation	[23,29]
477	F475	F490	E469	T481	<i>ApPDC</i> -E469G	inverts enantioselectivity for PAC towards (<i>S</i>)-PAC	[25]
					<i>BsMenD</i> -F490A/G	increased k_m and decreased activity for α -ketoglutarate and isochorismate	[16]
					<i>ECIPDC</i> -E468D	slightly decreased k_m but drastically reduced activity	[1]
					<i>EcAHASI</i> -V477I	increased k_m , decreased activity	[19]
					<i>EcGXC</i> -I479V	decreased activity in tartronate semialdehyde formation	[20]

continued Table 3.1: Variants of ThDP-dependent DCs at positions lining the S-pocket.

Std. pos. ^b	<i>EcMenD</i>	<i>BsMenD</i>	<i>ApPDC</i>	<i>PfBAL</i>	Variants ^c	Effect	References ^d
					<i>PpBFDC</i> - L461A/G/ S/V	variants do not affect substrate specificity; reduced activity and increased k_m in some variants (L461A,L461V); L461A inverts enantioselectivity towards (<i>S</i>)-acetoin; allows binding of <i>ortho</i> -substituted benzaldehyde derivatives	[21,23,26]
					<i>ScPDC</i> - E477Q	477 plays a role in the decarboxylation step; E477Q produces (<i>R</i>)-PAC and (<i>R</i>)-acetoin; E477Q increased activity in acetoin formation	[8,9,10,27]
					<i>ScPDC</i> - E477D/N/Q	mutations effect activity	[7,11,12]
					<i>ZmPDC</i> - E473A/D/I/ N/Q/T/V	decreased activity; E473Q binds ThDP more tightly than wt; E473I/V did not express in <i>E. coli</i> ; E473N/D diminished activity; E473Q converts <i>ZmPDC</i> from a decarboxylase into a carboligase	[15,28,29,30]
480	L478	L493	I472	F484	<i>BsMenD</i> - L493A	reduced k_m but also reduced activity for α -ketoglutarate and increased k_m /decreased activity for isochorismate	[16]
					<i>EcAHASI</i> - Q480W	reduced activity	[19]
					<i>EcAHASII</i> - W464L	variant still (<i>R</i>)-specific in PAC-formation; decreased sensitivity to product inhibition; influences acceptor specificity	[31,32]
					<i>LpPOX</i> - E483A/Q	reduced activity but k_m for pyruvate decreased; E483 interacts with carboxylic acid group of L-ThDP	[4]
					<i>NtAHASI</i> - W573I	conferes herbicide resistance	[33]

continued Table 3.1: Variants of ThDP-dependent DCs at positions lining the S-pocket.

Std. pos. ^b	<i>EcMenD</i>	<i>BsMenD</i>	<i>ApPDC</i>	<i>PfBAL</i>	Variants ^c	Effect	References ^d
					<i>PfBAL</i> -F484I	decreased ligase activity to benzoin, but remaining activity towards formation of (<i>R</i>)-2HPP	[5]
					<i>PpBFDC</i> -F464I	reduced activity with benzoylformate and inactive towards pyruvate; increased ee in formation of (<i>R</i>)-acetoin; enantioselectivity was inverted to (<i>R</i>)-2-HPP in double mutant A460I/F464I	[22,23,24]
					<i>ZmPDC</i> -I476A/E/F/L/V	I476E produced predominantly (<i>S</i>)-PAC; I476F decreases decarboxylase activity	[24,29]

^a The identical assignment of standard numbers to positions in the active sites of ThDP-dependent DCs was previously shown (Gocke 2007; Vogel, Widmann, et al. 2012; Westphal 2013), but modified to focus on the positions influencing the enantioselectivity.

^b Standard positions according to standard numbering scheme for ThDP-dependent DCs.

^c *ApPDC*, *Acetobacter pasteurianus* Pyruvate decarboxylase; *BsMenD*, *Bacillus subtilis* 2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase; *EcAHASI*, *Escherichia coli* Acetohydroxy-acid synthase I catalytic subunit; *EcAHASII*, *E. coli* Acetohydroxy-acid synthase II catalytic subunit; *EcGXC*, *E. coli* Glyoxylate carboligase; *ECIPDC*, *Enterobacter cloacae* Indole-3-pyruvate decarboxylase; *KdcA*, *Lactococcus lactis* branched-chain α -Ketoacid decarboxylase; *LpPOX*, *Lactobacillus plantarum* Pyruvate oxidase; *NtAHASI*, *Nicotiana tabacum* Acetohydroxy-acid synthase I catalytic subunit; *PfBAL*, *Pseudomonas fluorescens* Benzaldehyde lyase; *PpBFDC*, *Pseudomonas putida* Benzoylformate decarboxylase; *ScPDC*, *Saccharomyces cerevisiae* Pyruvate decarboxylase; *SvPPDC*, *Streptomyces viridochromogenes* Phosphonopyruvate decarboxylase; *ZmPDC*, *Zymomonas mobilis* Pyruvate decarboxylase

^d [1]Schütz, Golbik, et al. 2005, [2]Kneen, Pogozheva, et al. 2005, [3]Brandt, Kneen, et al. 2010, [4]Meyer 2009, [5]Janzen et al. 2006, [6]Polovnikova et al. 2003, [7]Kutter et al. 2009, [8]Liu et al. 2001, [9]Baykal et al. 2006, [10]Sergienko and Jordan 2001b, [11]Balakrishnan, Gao, et al. 2012, [12]Jordan, Nemeria, et al. 1998, [13]Johnen and Sprenger 2009, [14]Wu et al. 2000, [15]Chang, Nixon, and Duggleby 1999, [16]Dawson et al. 2010, [17]Le et al. 2004, [18]Yep, Kenyon, and McLeish 2006, [19]Belenky et al. 2012, [20]Nemeria, Binshtein, et al. 2012, [21]Gocke, Walter, et al. 2008, [22]Kara et al. 2011, [23]Siegert et al. 2005, [24]Yep and McLeish 2009, [25]Rother et al. 2011, [26]Lingen, Kolter-Jung, et al. 2003, [27]Jordan, Zhang, and Sergienko 2002, [28]Huang et al. 2001, [29]Pohl, Siegert, et al. 1998, [30]Meyer, Walter, et al. 2011, [31]Engel et al. 2003, [32]Tittmann, Vyazmensky, et al. 2005, [33]Chong et al. 1999

3.2.4 BioCatNet

The *BioCatNet* system developed as part of this work replaced the *DWARF* system as the in-house¹⁹ standard for the generation, maintenance and analysis of FSPDs. As already mentioned, the TEED²⁰, IRED²¹, CYPED²², TALED²³, LED²⁴, LacED²⁵, LccED²⁶, TTCED²⁷ and other FSPDs were recreated using *DBParse* and thereby incorporated into *BioCatNet* as of this writing. Moreover, due to the increased performance and the scalability, *BioCatNet* and the underlying software tools will facilitate prospective generation, maintenance and analysis of further FSPDs. The conceptual redesign from *DWARF* to *BioCatNet* further resulted in a combination of the formerly separated interfaces for database curation and database publication. This accelerates the process from the in-house database generation to the online publication. However, besides those arguments mainly relevant for those applicants using *BioCatNet* as a system for the generation of FSPDs, the renewal also affects scientists working in biocatalysis. Due to extension of the data model and development of an intuitive graphical user interface (GUI)²⁸ the *BioCatNet* system can cope with experimentally derived data. The mentioned GUI was designed to be accessible via the internet using any device with an up-to-date web browser in order to be suited as an electronic lab journal. Although it is not intended to be an alternative to classical lab journals, its application will help to ensure defined standards in the documentation of biocatalytic experiments. By demanding minimal requirements, this will in turn improve the comparability of the data generated by individual experimenters, the data generated within labs and the data shared between different working groups. In addition to this, the users profit from a well-formatted printout, which summarizes the submitted data, as well as from the possibility to

¹⁹Institute of Technical Biochemistry, University of Stuttgart, Germany

²⁰ThDP-dependent Enzyme Engineering Database (www.teed.biocatnet.de)

²¹Imine Reductase Engineering Database (www.ired.biocatnet.de). Database was generated and is maintained by Silvia Fademrecht, Institute of Technical Biochemistry, University of Stuttgart, Germany (Scheller et al. 2014)

²²Cytochrome P450 Engineering Database. The database was generated and maintained in collaboration with Łukasz Gricman, Institute of Technical Biochemistry, University of Stuttgart, Germany

²³TransALdolase Engineering Database. Database generated and maintained in collaboration with Lenz Lorenz, Institute of Technical Biochemistry, University of Stuttgart, Germany

²⁴Lipase Engineering Database. Database was generated and is maintained in collaboration with Silvia Fademrecht, Jennifer Häfner and Nora Schuth, Institute of Technical Biochemistry, University of Stuttgart, Germany

²⁵Lactamase Engineering Database. Database was generated and is maintained in collaboration with Catharina Zeil, Institute of Technical Biochemistry, University of Stuttgart, Germany

²⁶LaCCase and multicopper oxidase Engineering Database. Database was generated and is maintained in collaboration with Łukasz Gricman, Institute of Technical Biochemistry, University of Stuttgart, Germany

²⁷TriTerpene Cyclase Engineering Database. Database was generated and is maintained by Silvia Fademrecht, Institute of Technical Biochemistry, University of Stuttgart, Germany

²⁸The GUI was developed by Waldemar Reusch working as a research assistant and subsequently in the scope of his diploma thesis (2013-2014, Institute of Technical Biochemistry, University of Stuttgart, Germany).

browse released functional information in order to identify promising biocatalysts for desired reactions and to support the experimental design concerning appropriate reaction conditions. It is noteworthy that *BioCatNet* obligates the users to define product concentrations. However, since it allows declaration of 0 mol/L as the product concentration it can also be used to document and identify substrates and reaction conditions that do not work for conversions with desired enzymes. Consequently, *BioCatNet* has the potential to reduce cost and effort in biocatalysis. In addition, the system can provide a well-formatted, easily accessible and highly interlinked long-term storage of biochemical data, which is claimed by various research funding organisations. On the other hand, bioinformatical analysis will likewise benefit from the use of *BioCatNet* in laboratories since comparable and easily accessible data can be automatically analyzed. By the required linkage of submitted biochemical data with the sequence of the respective biocatalysts, *BioCatNet* is intended to overcome the limitations of literature mining, where unambiguous information about the enzymes' amino acid sequence often is lacking. Moreover, implementation of standard numbering schemes makes comparison of different enzymes and the effects of different variants feasible without need for structure information. Thus, the *BioCatNet* system provides promising possibilities to identify specific features of different enzymes and variants that effect the functionality - by a systematic analysis of the sequence-structure-function relationships of enzymes. The rational engineering done as part of this work proved the potential of FSPDs to support protein design. With *BioCatNet*, this potential even grows further.

3.3 Conclusions and future perspectives

For a systematic analysis of the sequence-function relationships, biochemical information unambiguously linked to amino acid sequences must be available for different enzymes and enzyme variants. Moreover, the data derived from biocatalytic experiments must be comparable in respect to the respective reaction conditions. Further, the information must be accessible in a format that can be used for computational analysis. Literature mining approaches were used by others to collect biochemical information but the accuracy, completeness, comparability and unambiguous link to the enzymes' sequences often remain doubtful or are missing. In order to overcome the limitations of literature mining, which are actually caused by missing or insufficiently strict requirements by the journals publishing scientific results, a system to collect accurate information directly at the experimenter's bench would be desirable. By claiming minimal requirements to assure comparability of the submitted data and by linking all experimental results with the amino acid sequences of the respective biocatalysts, systematic analysis of the sequence-function and structure-function relationships would be enabled. So far, there is no such comprehensive repository on biochemical information about ThDP-dependent enzymes available. The *BioCatNet* system is intended to close this gap and to enable systematic large-scale analysis. However, since the development of the *BioCatNet* system just recently reached a state allowing data acquisition, there is not sufficient experimentally derived information about biocatalytic results included to enable a systematic analysis yet. Consequently, the systematic analysis of the sequence-function or structure-function relationships remains open, due to the lack of computationally accessible and sequence-associated biochemical data of enzymes. Nevertheless, the work presented in this thesis laid the foundation for a systematic acquisition of accurate experimental data and provides the tools needed for future analysis of those relationships.

In addition, more focused analysis of the sequences and structures of selected ThDP-dependent DCs enabled rational engineering and revealed parts of the sequence-function and structure-function relationships of ThDP-dependent enzymes. Robert Westphal confirmed the *S*-pocket concept proposed for ThDP-dependent enzymes by engineering *EcMenD* and *BsMenD* (Westphal, Waltzer, et al. 2013; Westphal, Hahn, et al. 2013; Westphal, Jansen, et al. 2014; Westphal 2013). Based on predictions of the *S*-pocket contributing residues using the standard numbering scheme for ThDP-dependent DCs and investigation of the available crystal structures, the hotspots determining the enantioselectivity were readily identified and mutated. Moreover, application

of the standard numbering scheme for ThDP-dependent DCs, subsequent analysis of the amino acid distribution in enzymes homologous to *ApPDC* and *PfBAL*, and molecular modeling of putative substrate orientations in both active sites, enabled prediction and generation of an enzyme capable to catalyze the desired direct asymmetric synthesis of (*S*)-benzoins (Westphal et al. 2014a).

A systematic approach was further applied to use the sequence-structure relationship of ThDP-dependent enzymes for the generation of homology models. Implemented in the TEED and enriched with information needed for quality estimation, those models can now be used for a systematic analysis of the (putative) structures of ThDP-dependent enzymes. To conclude, the sequence-structure, sequence-function and structure-function relationships for ThDP-dependent enzymes are at least partially understood. First investigations of those relationships in transketolases revealed similar principles but did also show differences as compared to the more intensively investigated DCs. The region in transketolases corresponding to the *S*-pocket in members of the DC superfamily does not provide the space needed for an antiparallel substrate arrangement. However, identification of the corresponding residues using the standard numbering scheme for the PP and PYR domains of ThDP-dependent enzymes and *in silico* analysis of available crystal structures enabled prediction of a variant of *EcTK* with a pocket corresponding to the *S*-pocket of DCs. As part of her doctorate, Anna Baierl²⁹, is investigating the enantioselectivity determining factors in transketolases and is going to introduce this pocket. Thus, although there is indication of a different regulation of the enantioselectivity in transketolases, transplantation of this pocket might broaden the substrate spectrum towards benzaldehyde as acceptor substrate and might introduce an additional and well-understood switch for the enantioselectivity in transketolases.

To conclude I would like to refer back to the quote from David Lipman about Margaret Dayhoff, which introduced this thesis. Bioinformatics comprises three major components: biological discoveries, tool development and resource development. In this work all three components were combined in order to facilitate the systematic analysis of the sequence-structure-function relationships and in kind collaboration with other scientists, the 'biological discoveries' were used to successfully perform rational engineering.

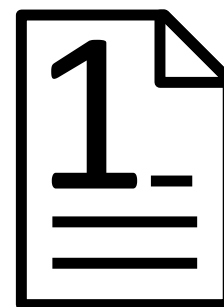
²⁹Institute of Bio- and Geosciences (IBG-1), Biotechnology, Forschungszentrum Jülich GmbH, Germany

4 Publications

Parts of this thesis have already been published during the last years. For a complete list of publications see page 229. In the following, four publications playing a particular role in this thesis are appended, since they provide additional information on the methodical details and further results.

4.1 A standard numbering scheme for thiamine diphosphate-dependent decarboxylases

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



4.1.1 Abstract

Background: Standard numbering schemes for families of homologous proteins allow for the unambiguous identification of functionally and structurally relevant residues, to communicate results on mutations, and to systematically analyze sequence-function relationships in protein families. Standard numbering schemes have been successfully implemented for several protein families, including lactamases and antibodies, whereas a numbering scheme for the structural family of thiamine diphosphate (ThDP)-dependent decarboxylases, a large subfamily of the class of ThDP-dependent enzymes encompassing pyruvate-, benzoylformate-, 2-oxo acid-, indolpyruvate- and phenylpyruvate decarboxylases, benzaldehyde lyase, acetohydroxyacid synthases and 2-succinyl-5-enolpyruvyl-6- hydroxy-3-cyclohexadiene-1-carboxylate synthase

(MenD) is still missing. Despite a high structural similarity between the members of the ThDP-dependent decarboxylases, their sequences are diverse and make a pairwise sequence comparison of protein family members difficult.

Results: We developed and validated a standard numbering scheme for the family of ThDP-dependent decarboxylases. A profile hidden Markov model (HMM) was created using a set of representative sequences from the family of ThDP-dependent decarboxylases. The pyruvate decarboxylase from *S. cerevisiae* (PDB: 2VK8) was chosen as a reference because it is a well characterized enzyme. The crystal structure with the PDB identifier 2VK8 encompasses the structure of the *ScPDC* mutant E477Q, the cofactors ThDP and Mg^{2+} as well as the substrate analog (2S)-2-hydroxypropanoic acid. The absolute numbering of this reference sequence was transferred to all members of the ThDP-dependent decarboxylase protein family. Subsequently, the numbering scheme was integrated into the already established Thiamine diphosphate-dependent Enzyme Engineering Database (TEED) and was used to systematically analyze functionally and structurally relevant positions in the superfamily of ThDP-dependent decarboxylases.

Conclusions: The numbering scheme serves as a tool for the reliable sequence alignment of ThDP-dependent decarboxylases and the unambiguous identification and communication of corresponding positions. Thus, it is the basis for the systematic and automated analysis of sequence-encoded properties such as structural and functional relevance of amino acid positions, because the analysis of conserved positions, the identification of correlated mutations and the determination of subfamily specific amino acid distributions depend on reliable multisequence alignments and the unambiguous identification of the alignment columns. The method is reliable and robust and can easily be adapted to further protein families.

4.1.2 Background

Thiamine diphosphate (ThDP) -dependent decarboxylases are a large subfamily of the class of ThDP-dependent enzymes which are essential in many biosynthetic pathways. Due to the scientific and industrial relevance of enzymes capable of catalyzing C-C bond formation and cleavage, we have focused in this work on the decarboxylase superfamily of the ThDP-dependent Enzyme Engineering Database (TEED) (Widmann, Radloff, and Pleiss 2010). This superfamily

contains among others pyruvate decarboxylases (PDCs, EC 4.1.1.1), indolepyruvate decarboxylases (IPDCs, EC 4.1.1.74), pyruvate oxidases (POXs, EC 1.2.3.3), pyruvate dehydrogenases (PDHs, EC 1.2.4.1), oxalyl-CoA decarboxylases (OCDCs, EC 4.1.1.8), benzaldehyde lyases (BALs, EC 4.1.2.38), benzoylformate decarboxylases (BFDs, EC 4.1.1.7), acetohydroxyacid synthases (AHASs, EC 2.2.1.6), glyoxylate carboligases (GXC, EC 4.1.1.47), sulfoacetaldehyde acetyltransferases (SAATs, EC 2.3.3.15), 2-hydroxyphytanoyl-CoA lyases (2-HPCLs) and 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexadiene-1-carboxylate synthases (SEPHCHC, MenD).

Despite low sequence similarities between sequences of the decarboxylase superfamily of the TEED (~ 20%), their structures are highly similar. The structures consist of three domains, the N- and C-terminal domains are involved in binding of the cofactor ThDP and are named pyrimidine (PYR) and pyrophosphate (PP) binding domain (Duggleby 2006; Costelloe, Ward, and Dalby 2008), respectively. They are separated by a third domain, which is less conserved and adopts different functions in the various enzyme families, e.g. by binding additional cofactors such as ADP (Werther et al. 2010) and FAD (Lindqvist and Schneider 1993) or activators and inhibitors (Kutter et al. 2009). Due to structural relations between this middle domain and the transhydrogenase domain dIII, this domain is called the TH3 domain (Duggleby 2006; Costelloe, Ward, and Dalby 2008).

Although all ThDP-dependent decarboxylases share the same fold and a similar mechanism utilizing the cofactor ThDP, they catalyze a broad range of different reactions involving cleavage and formation of C-C bonds (Pohl, Sprenger, and Müller 2004; Pohl, Lingen, and Müller 2002; Müller, Gocke, and Pohl 2009). While the decarboxylation of 2-ketoacids (Kluger and Tittmann 2008) and the carboligation of two aldehydes to 2-hydroxy ketones are catalyzed by most members of the ThDP-dependent decarboxylases (Müller, Gocke, and Pohl 2009), their substrate ranges are different. The well characterised PDC from *Saccharomyces cerevisiae*, BFDC from *Pseudomonas putida* and BAL from *Pseudomonas fluorescence* accept a broad variety of substrates (Pohl, Sprenger, and Müller 2004; Pohl, Dresen, et al. 2012; Pohl, Gocke, and Müller 2010), while SEPHCHC-synthase (MenD) is limited to a small number of substrates (Müller, Kurutsch, et al. 2009; Fang, Macova, et al. 2011).

Additional complexity of C-C bond formation results from the fact that a substrate might be either a donor, which is activated by addition to ThDP in the active site, or an acceptor, which reacts

with the ThDP-bound donor, resulting in different products (Pohl, Sprenger, and Müller 2004; Pohl, Dresen, et al. 2012; Pohl, Gocke, and Müller 2010). Reactions catalyzed by members of the structural group of ThDP-dependent decarboxylases include decarboxylation of 2-keto acids, synthesis of various chiral 2-hydroxy ketones by asymmetric benzoin- (Pohl, Dresen, et al. 2012; Demir, Sesenoglu, Dünkermann, et al. 2003) and cross-benzoin condensation (Dünkermann et al. 2002; Dünwald and Müller 2000), the racemic resolution of 2-hydroxy ketones via C-C bond cleavage (Demir, Pohl, et al. 2001), and Stetter-like reactions, e.g. the addition of decarboxylated 2-ketoglutarate to isochorismate by MenD (Jiang et al. 2007).

With the exception of a few functionally relevant residues that have been identified by comparing sequences and structures of homologous proteins or by mutation experiments, the molecular basis of this biochemical diversity is still unknown. Variants have been developed by rational design and by directed evolution in order to improve the activity of members of this enzyme family (Dünkermann et al. 2002; Jordan, Li, and Brown 1999; Li, Furey, and Jordan 1999) or to alter substrate specificity (Pohl, Siegert, et al. 1998; Liu et al. 2001; Sergienko and Jordan 2001b; Sergienko and Jordan 2001a; Siegert et al. 2005; Lingen, Kolter-Jung, et al. 2003; Andrews and McLeish 2012) or stereoselectivity (Rother et al. 2011; Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008). Some functionally relevant amino acids are located in the active site, mediating substrate binding (Costelloe, Ward, and Dalby 2008), are involved in the activation of ThDP (Andrews and McLeish 2012) or steer stereoselectivity (Rother et al. 2011; Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008), e.g. the *S*-pocket as part of the acceptor binding site, which has been shown to contribute to the stereoselectivity of several members of the decarboxylase superfamily (Rother et al. 2011; Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008).

However, due to this complexity, combining results yielded from different variants of different protein families, consolidating results on the function of specific residues and comparing results from different research groups is unfortunately not a straightforward process. An additional challenge in this respect is the identification of homologous positions in sequences of different proteins in order to allow for their comparison. Amino acid exchanges in enzyme variants are usually identified by a number, signifying the absolute position of the amino acid in the respective protein in combination with the original and the newly introduced amino acid. This method only yields comparable results if the numbering is based on exactly the same sequence. In reality however, published results often are based on slightly different protein sequences, often missing residues at the N-terminus or based on sequences derived from crystal structures. This makes

the comparison of results concerning individual residues of one specific protein from different research groups or the comparison of results on homologous proteins manually intensive and prevents the use of automated tools for a large number of sequences. Therefore, an unambiguous numbering scheme for all members of the decarboxylase superfamily would be desirable.

The usefulness of a generally accepted numbering scheme was demonstrated for the class A and B enzyme families of β -lactamases (Galleni et al. 2001; Ambler 1980). Based on structure-guided multisequence alignments of reference sequences (Garau et al. 2004), a number was assigned to each column of the alignment. Thus, each amino acid could be addressed unambiguously and consistently for all sequences. This numbering scheme is widely applied for the identification of key residues and for the naming of variants (Garau et al. 2004). The numbers assigned by this scheme might differ by more than 20 from the absolute amino acid numbering of a respective protein. Without a standard numbering scheme the systematic comparison of mutations would have to be done manually and would be error-prone. For the same reasons, a standard numbering scheme was established for complementary-determining regions (CDRs) of antibodies, thus allowing for a systematic analysis and an unambiguous communication between research groups (Al-Lazikani, Lesk, and Chothia 1997; Kabat, Wu, and Perry 1991). The numbering schemes were initially based on limited sets of protein sequences and were subsequently refined as more sequence and structure data became available. In order to provide a standard numbering which is independent from the increasing sequence space, a numbering scheme based on one defined reference sequence would be desirable. Due to the low sequence similarity between ThDP-dependent decarboxylases from different homologous families, it would not be reliable to transfer the absolute position numbers of the reference sequence to the residues of any decarboxylase sequence based on pairwise alignments. To handle this challenge, we chose a structure-based and profile-guided approach for the transfer of position numbers. In this work, we present the establishment of a numbering scheme for the ThDP-dependent decarboxylases based on the sequence of the well-documented pyruvate decarboxylase from *S. cerevisiae* (pdb|2VK8, Kutter et al. 2009, sp|P06169). The numbering scheme was validated by comparing its ability to produce multisequence alignments to the T-Coffee alignment algorithm and by revision of the structural equivalence of positions with the same standard numbers. Using this numbering scheme, the decarboxylase superfamily was systematically analyzed for conserved amino acids.

4.1.3 Results

Implementation and validation of a standard numbering scheme

A standard numbering scheme for the decarboxylase superfamily of ThDP-dependent enzymes was established using the ThDP-dependent Enzyme Engineering Database (TEED) (Widmann, Radloff, and Pleiss 2010). A profile hidden Markov model was created from a structure-guided multisequence alignment of 16 representative proteins of the decarboxylase superfamily (Table 4.1). One of the representative proteins, the pyruvate decarboxylase from *S. cerevisiae* (ScPDC, sp|P06169, pdb|2VK8, Kutter et al. 2009), was used as the reference sequence for numbering all proteins of the decarboxylase superfamily. In addition, 22 functionally and structurally relevant residues in the sequence of ScPDC were annotated as described in literature (Duggleby 2006; Lindqvist and Schneider 1993; Andrews and McLeish 2012; Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008; Takenaka et al. 2007; Lobell and Crout 1996; Jordan 2003) (Supplementary Information, Table A.3 on page 190).

Table 4.1: The set of 16 representative proteins used for establishing a standard numbering scheme. Of each PDB entry, chain A was used for the alignment. It was verified that for all proteins chain A corresponds to the catalytic subunit.

Protein	Organism	PDB-identifier
pyruvate decarboxylase	<i>S. cerevisiae</i>	2VK8
2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexadiene-1-carboxylate synthase	<i>E. coli</i>	2JLC
pyruvate decarboxylase	<i>Z. mobilis</i>	1ZPD
branched-chain keto acid decarboxylase	<i>L. lactis</i>	2VBF
benzoylformate decarboxylase	<i>P. putida</i>	1BFD
carboxyethylarginine synthase	<i>S. clavuligerus</i>	2IHT
cyclohexene-1,2-dione hydrolase	<i>Azoarcus sp.</i>	2PGN
oxalyl-CoA decarboxylase	<i>O. formigenes</i>	2C31
pyruvate oxidase	<i>A. viridans</i>	1V5F
pyruvate dehydrogenase	<i>E. coli</i>	3EYA
indolepyruvate decarboxylase	<i>E. cloacae</i>	1OVM
acetohydroxyacid synthase	<i>S. cerevisiae</i>	1JSC
acetohydroxyacid synthase	<i>A. thaliana</i>	1YBH
acetohydroxyacid synthase	<i>K. pneumoniae</i>	1OZF
benzaldehyde lyase	<i>P. fluorescens</i>	2AG0
glyoxylate carboligase	<i>E. coli</i>	2PAN

These positions include the highly conserved active site residues E51 (standard numbering) (Candy, Koga, et al. 1996; Schütz, Sandalova, et al. 2003; Killenberg-Jabs et al. 1997; Shaanan

and Chipman 2009), the conserved HH motif in PDCs (H114/H115) (Andrews and McLeish 2012), the GDGX motif 443-446 and the Mg²⁺ binding site N471 (Jordan 2003), as well as more variable regions such as the S-pocket residues P26, G27, I476, and Q477 (Rother et al. 2011; Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008) and the start and end position of the three decarboxylase domains, the PYR, PP, and the TH3 domain (Duggleby 2006). In contrast to the PYR and the PP domain, the secondary structure elements of the TH3 domains of different decarboxylases vary considerably near their N- and C-terminus, thus leading to numerous gaps in the alignment at these positions. Therefore, the start of the TH3 domain was shifted four positions downstream and the end was shifted five positions upstream into regions, which were free of gaps, though sequence conservation was still low. The absolute amino acid numbers and annotation information were transferred from the reference sequence to the respective positions of all members of the decarboxylase superfamily by aligning them to the profile HMM. A web application was integrated into the web interface of the TEED (www.TEED.uni-stuttgart.de) to provide public access to the numbering tool. Upon submission of a single query sequence or a list of sequences in FASTA format, the standard numbering is applied and the sequence including the numbering and annotations for each amino acid can be downloaded (Figure 4.1 on the next page; a description of the file format and a sample are given in the Supplementary Information, A.4.1 on page 194 and A.1 on page 194).

The accuracy of the HMM-based alignment was compared to a multisequence alignment using T-Coffee (Notredame, Higgins, and Heringa 2000) by aligning the reference sequence *ScPDC* and 15 sequences from the decarboxylase family for which structural information was available but which were not part of the set of representative proteins. In order to determine the differences between the HMM-based alignment and the T-Coffee alignment, all columns were compared between the two alignments and a similarity score was assigned to each column (Supplementary Information, Section A.4.3 on pages 197ff.). Alignment columns were 'identical' if both alignment algorithms placed the same residues for all sequences into the respective columns; 'highly similar' if the two alignments differed in 1-3 sequences; 'similar' if 4-8 mismatches were observed; 'dissimilar' if 9 - 12 sequences differed at the respective position; 'divergent' if the alignments differed in 13 - 15 of the 15 sequences. As a result, 73% of all columns were identical or highly similar in both alignments (Figure 4.2 on page 84). For those columns which deviated considerably between the two alignments (dissimilar or divergent columns), a structural comparison revealed that in almost all cases the HMM-based alignment represented the structural

query:	M---YTVGDY	LLDRLHELGI	EEIFGVPGDY	NLQFLDQIIS	REDMKWIGNA	47
reference:	-MSEITLTKY	LFERLKQVNV	NTVFLPLGDF	NLSLLDKIYE	VEGMRWAGNA	49
query:	NELNASYMAD	GYARTKAAA	FLTTFGVVEL	SAINGLAGSY	AENLPVVEIV	97
reference:	NELNAAYAAD	GYARIKMSC	IITTFGVVEL	SALNGIAGSY	AEHVGVLVHV	99
query:	GSPTSKVQND	GKFBVHHTLAD	GDFKHFMMKH	EPVTAARTLL	TA-ENATYEI	146
reference:	GVPSISAQAK	QLLHHTLGN	GDFTVFHRMS	ANISETTAMI	TDIATAPAEI	149
query:	DRVLSQLLKE	RKPVYINLEV	DVAAAKAEKP	ALSLEKESST	-TNT---TEQ	192
reference:	DRCIRTTYVT	QRPVYLGLPA	LLQ-TPIDMS	LKPNDAESEK		198
			PYR-end standard: 168 query: 165			
query:	VILSKIEESL	KNAQKPVVIA	KTVTQFVSET	KLPITTLNFG		242
reference:	EVIDTILVLD	KDAKNPVILA	DACCSRHDVK	AETKKLIDLT	QFPFVTPMG	248
query:	KSAVDESPLS	FLGIYNGKLS	EISLKNFVES	ADFILMLGVK	LTDSSSTGAFT	292
reference:	KGSIDEQHPR	YGGVYVGTLS	KPEVKEAVES	ADLILSVGAL	LSDFNTGSFS	298
query:	HHLDENKMIS	LNIDEGIIFN	KVVEDDFDRA	VVSSLSEL-K	G--IEYEGQY	339
reference:	YSYKTKNIVE	FHSDHMKIRN	ATFPGVQMKF	VLQKLLTIA	DAAKGYK-P-	346
query:	IDKQ---YE-	E--FIPSSAP	LSQDRLWQAV	ESLTQSNETI	VAEQGTSFFG	383
reference:	-VAVPARTPA	NAA-VPASTP	LKEWMMWNQL	GNFLQEGDVV	IAETGTSAFG	394
query:	ASTIFLKSNS	RFIGQPLWGS	IGYTFPAALG	SQIAD----K	ESRHLLFIGD	429
reference:	INQTTFPNNT	YGISQVLWGS	IGFTTGATLG	AAFAAEEIDP	KKRVLFIGD	444
query:	GSLQLTVQEL	GLSIREKLN	ICFIINNDGY	TVEREIHGPT	QSYNDIPMWN	479
reference:	GSLQLTVQEI	STMIRWGLKP	YLFVLNNDGY	TIQKLIHGPK	AQYNEIQGWD	494
query:	YSKLPETFGA	TEDRVVSKIV	RTENEFVSVM	KE--AQADV	RMYWIELVLE	527
reference:	HLSLLPTFG-	AKD-YETHRV	ATTGEWDKLT	QDKSFN-DNS	KIRMIEVMLP	541
query:	KEDAPK----	---LLKKMGK	LFAEQNK			
reference:	VFDAPQNLVE	QAKLTAATNA	KQ-----			

Figure 4.1: Alignment of a query sequence and the reference sequence from the web interface of the numbering method. Alignment of a query sequence (here: branched-chain alpha-ketoacid decarboxylase from *L. lactis*, accession number: gi|75369656) to the reference sequence (*ScPDC*, accession number: sp|P06169, pdb|2VK8). By positioning the cursor on an amino acid (here: proline), the standard numbering (here: 168) as derived from the reference sequence and the absolute numbering of the respective query sequence (here: 165) as well as annotation information (here: end of the PYR domain) are displayed. All residues are highlighted for which annotation information is available in the TEED (Widmann, Radloff, and Pleiss 2010).

equivalence better than the multisequence alignment by T-Coffee (Supplementary Information, Figures A.3 on page 189 and A.2 on page 188). In addition, it was verified that all 22 functionally relevant positions were aligned correctly (Supplementary Information, Table A.3 on page 190).

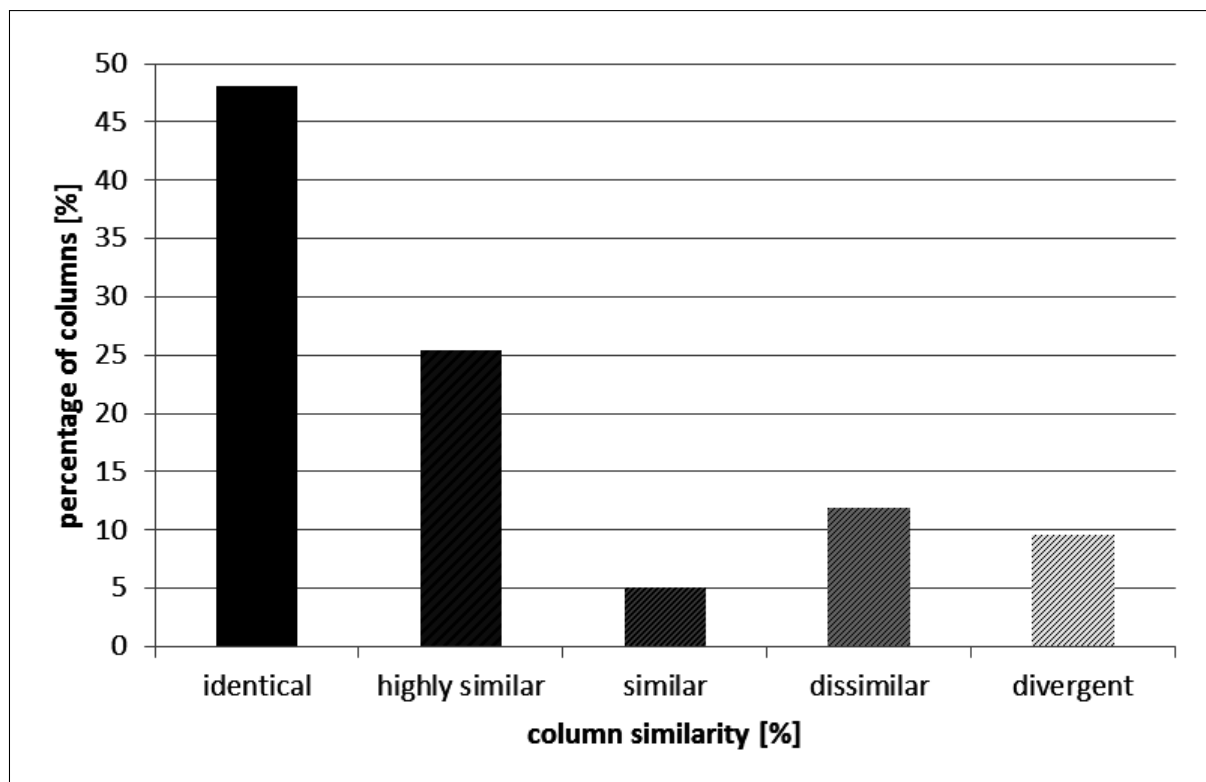


Figure 4.2: Analysis of accordance of two multisequence alignments. The comparison of columns of two multisequence alignments of 15 sequences using the numbering method and T-Coffee revealed five types of column similarity. 48% of the investigated columns were identical in both alignments, 25% of the columns were 'highly similar' (up to 3 mismatches out of 15 sequences), 5% were 'similar' (4 - 8 mismatches), 12% of the columns had 9 to 12 mismatches and are therefore called 'dissimilar' and 10% of the columns showed more than 12 mismatches ('divergent').

Identification of conserved residues and domain boundaries

After having applied a standard numbering scheme for all 3000 members of the decarboxylase superfamily, the respective protein sequences were systematically analyzed for the occurrence of amino acids at corresponding positions. Four groups of positions with different characteristics of conservation were found.

The first group includes 6 positions which were conserved in more than 90% of all members of the decarboxylase superfamily, while no other amino acid occurred in more than 1% of the sequences: Position 27 (standard numbering) in the S-pocket which was glycine in 91% of all members of the decarboxylase superfamily, position 443 in the GDGX motif which was glycine in 98% of all decarboxylases, and four highly conserved positions which have not been identified yet as being of functional or structural relevance: positions 58 (alanine in 96% of the sequences), 94 (proline in 91% of the sequences), 219 (glycine in 91% of the sequences), and 286 (glycine in 97% of the sequences). Thus, 6 positions (mostly glycine residues) are highly conserved in

almost all members of the decarboxylase superfamily.

The second group includes 3 positions in which one amino acid was found in a majority of more than 90% of all members of the decarboxylase superfamily and a different amino acid in a minority (> 3%) of all sequences. The most conserved position was the active site residue Glu 51. This conserved glutamic acid was found in 94% of all sequences, while 3% have a valine in this position. D444 of the GDGX motif was conserved in 91% of all cases, while 7% have a glutamic acid in this position. At position 280, aspartic and glutamic acid were found in 90% and 4%, respectively, of all members of the decarboxylase superfamily. Thus, this group includes positions which seem to be characteristic for a distinct subgroup of this superfamily.

The third group encompasses variable positions which are known to be involved in substrate recognition or catalysis. In positions 114 and 115, the majority of all members of the decarboxylase superfamily have a phenylalanine (58%) and a glutamine (81%), respectively, while a minority, predominantly PDCs, show histidine (15% and 12%, respectively) in these positions. These histidines have been referred to as the HH-motif in the PDC family (Andrews and McLeish 2012). A functionally relevant, though highly variable site, is the *S*-pocket which contributes to the stereo selectivity of decarboxylases (Rother et al. 2011; Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008). Two positions, 476 and 477, which were shown to contribute to the *S*-pocket or the entrance of the *S*-pocket, were highly variable in all members of the decarboxylase superfamily. In standard position 476 most members of the decarboxylase superfamily show a methionine (42%) or an isoleucine residue (18%), respectively, while standard position 477 is occupied by valine (45%) or isoleucine (20%), respectively.

The fourth group included the domain boundaries of the three protein domains PYR, PP and the TH3 domain. Identification of the domain boundaries can be easily accomplished when structural information is available, whereas an identification of domain boundaries based on the amino acid sequence alone is not straightforward due to the low sequence similarity in the loop regions connecting the three domains. However, alignments using the profile HMM revealed several conserved positions: the start of the PYR domain (standard numbering 6) is indicated by a conserved glycine (in 44% of all sequences), while its end (position 168) is highly conserved (proline in 87% of all cases). Similarly, the PP domain starts at position 367 (proline in 54% of all sequences) and ends at position 540 (valine in 37% of all sequences). These four positions coincided well with the start and end of the ThDP-binding fold. In contrast, the start and end

positions of the TH3 domain were highly variable. Therefore, two positions further inside the TH3 domain were selected to characterize the start and the end of this domain: positions 197 (aspartic acid in 18% of all cases) and 336 (lysine in 17% of all sequences). Despite the low sequence similarity in the boundary region, the assignment of standard numbers was consistent with the results from a structural superimposition.

Furthermore, the regions around the 9 highly conserved positions of group 1 and 2 were investigated concerning sequence conservation in order to investigate the presence of sequence motifs. With the exception of position 27 (standard numbering) their surrounding regions were sufficiently conserved to allow for the derivation of sequence motifs. The region around residue G443 is already known as the GDGX_{24,27}N-motif (Hawkins, Borges, and Perham 1989). In order to analyze the specificity and the precision of the remaining motifs for the decarboxylase superfamily, they were used in a motif search against the non-redundant NCBI database, while an updated version of the TEED (not yet published) served as positive control (Widmann, Radloff, and Pleiss 2010). The motif [DHN]₅₀-E₅₁-[A EGLQ]₅₂-[AGNSTV]₅₃-[AGLMV]₅₄-[AGISTV]₅₅-[FHLMY]₅₆-[AFILM]₅₇-A₅₈, which was derived from the region around the conserved positions 51 and 58, showed similar sensitivity (0.65) and precision (0.27) as the PROSITE pattern PS00187, which is an extended version of the GDGX_{24,27}N-motif and was described as a conserved motif of POXs (EC 1.2.3.3), PDCs (EC 4.1.1.1), AHASs (EC 2.2.1.6), BFDCs and indolepyruvate decarboxylases (IPDCs, EC 4.1.1.74) (Green 1989; Koga, Adachi, and Hidaka 1991; Tsou et al. 1990) (sensitivity: 0.59, precision: 0.42). This motif is part of an α -helix, which is involved in the formation of the active site. In addition, the motif surrounding position 280 had at least similar precision and sensitivity for ThDP-dependent decarboxylases as the simple GDGX_{24,27}N-motif (Hawkins, Borges, and Perham 1989) (data not shown). Thus, a second motif [DE]₂₈₀-[ACFLTV]₂₈₁-[ILMV]₂₈₂-[FILV]₂₈₃-[ACGLMNSTV]₂₈₄-[AFILV]₂₈₅-G₂₈₆ was identified with 5 predominantly hydrophobic amino acids between two highly conserved positions D/E280 and G286, which form the vertices of the loops connecting a central β -strand of the TH3 domain to the adjacent α -helices. The remaining motifs were less specific and sensitive for the identification of ThDP-dependent decarboxylases.

Application of the numbering scheme to experimentally characterized positions

An extensive literature search yielded 22 positions, which were experimentally well characterized in five different proteins (*ScPDC*, *ApPDC*, *ZmPDC*, *PfBAL* and *PpBFDC*) and shown to be of relevance to substrate specificity and/or activity. The numbering scheme was exemplarily applied to the respective sequences in order to compare the annotation information from the literature. Several equivalent positions in different proteins were shown to have different absolute numbers (Supplementary Information, Table A.3 on page 190). An influence on the decarboxylase activity was shown for the residues D28 of *ScPDC* (Liu et al. 2001; Sergienko and Jordan 2001b; Sergienko and Jordan 2001a), D27 of *ZmPDC* (Chang, Nixon, and Duggleby 1999; Huang et al. 2001; Wu et al. 2000) and A28 of *PfBAL* (Brandt, Kneen, et al. 2010; Brandt, Nemeria, et al. 2008; Janzen et al. 2006; Kneen, Pogozeva, et al. 2005), each corresponding to standard position 28. Furthermore, structural and functional equivalence was shown for A28 in *PfBAL* and S26 in *PpBFDC*. Similarly, positions 114 and 115, which were described as the HH-motif of pyruvate decarboxylases (Andrews and McLeish 2012) are structurally and functionally identical in different PDCs, but differ in their absolute position numbers (Supplementary Information, Table A.4 on page 191). The mutations W388A,I in *ApPDC* were shown to reduce stereoselectivity (Rother et al. 2011) while the mutations W392A,I,M of *ZmPDC* led to an improved carboligation activity (Bruhn et al. 1995; Goetz et al. 2001; Iwan et al. 2001; Pohl 1997). However, both positions are structurally equivalent and are addressed with standard number 392. Functional relevance is also described for position 477 (standard number) in *ScPDC*, *ApPDC* and *ZmPDC*. All mutations of the respective residues (E477Q in *ScPDC* (Liu et al. 2001; Sergienko and Jordan 2001b; Sergienko and Jordan 2001a), E469G in *ApPDC* (Rother et al. 2011) and E473D,Q in *ZmPDC* (Chang, Nixon, and Duggleby 1999; Huang et al. 2001; Breslow 1957; Meyer, Neumann, Parthier, et al. 2010; Meyer, Walter, et al. 2011)) revealed an impact on the decarboxylation reaction. The examination of these five examples and the differences between the absolute and the standard numbers of functionally equivalent positions showed, that the presented numbering scheme for the ThDP-dependent decarboxylases facilitates the communication on variants and the comparison of functionally relevant positions. The assignment of standard numbers to positions of different homologous proteins furthermore simplifies the prediction of the impact of mutations at equivalent positions.

4.1.4 Discussion

A standard numbering scheme has been established for the structural superfamily of ThDP-dependent decarboxylases, as it has been done previously for two protein families, the β -lactamase family and the complementary determining regions of antibodies (Garau et al. 2004; Al-Lazikani, Lesk, and Chothia 1997). A standard numbering scheme for a protein family enables an unambiguous communication between research groups about corresponding positions in different proteins and supports the automated systematic analysis of sequences and the classification of proteins into sub-groups (Widmann, Pleiss, and Oelschlaeger 2012). In principle, a numbering scheme could be established by performing pairwise alignments of each sequence of the protein family to a reference sequence. However, although structurally conserved, the superfamily of ThDP-dependent decarboxylases shows only low sequence similarity. As a consequence, pairwise alignments are in general not reliable.

As an alternative, multisequence alignment methods were successfully applied to align homologous proteins with low sequence similarity (Thompson et al. 2011). By performing a multisequence alignment of all sequences of the decarboxylase superfamily, the numbering of a reference sequence could be transferred to each aligned decarboxylase sequence. However, a new alignment has to be calculated for each new sequence to be included. Calculating multisequence alignments of many thousands of sequences with low sequence similarity are not only computationally intensive, but more importantly, they lack robustness, because the alignment might change upon inclusion of additional sequences. In contrast, profile hidden Markov models (HMM) based on a structure-driven alignment are a robust description of protein families and allow the user to align new sequences to an existing multisequence alignment (Sievers et al. 2011). By alignment of a sequence to a profile built from a set of representative proteins, the numbering can be transferred from the reference sequence to a query sequence.

However, the quality of the numbering depends on the quality of the profile. Therefore, the proteins in the profile HMM were carefully selected. From each of the sixteen families with structural information, a representative protein was selected for a structure-guided alignment (Russell and Barton 1992) to guarantee the structural equivalence in the reference profile. Because some members of the decarboxylase superfamily show activation upon binding of a substrate at a second (allosteric) binding site (e.g. *ScPDC*) (Kutter et al. 2009) which leads to conformational changes, the set of reference proteins only contained decarboxylases, which show no substrate

activation or which have been crystallized in complex with an allosteric activator. Thus, only structures of active enzymes were compared. The alignment was further manually refined in order to improve consistency and robustness.

Since the presented numbering scheme is aimed to compare structurally equivalent positions, the method depends on structural similarity of the proteins in the corresponding family. Accordingly, the method can be adapted to other protein families matching this requirement globally or at least in structurally conserved domains. By establishing a standard numbering scheme for the ThDP-dependent decarboxylase superfamily, the unambiguous identification, numbering, and analysis of functionally and structurally relevant residues was possible. The analysis of conserved positions in the protein family of ThDP-dependent decarboxylases revealed that the previously observed substitution of the active site glutamate by valine in members of the glyoxylate carboligase family at standard position 51 (Shaanan and Chipman 2009; Kaplun et al. 2008) is indeed characteristic of the entire family, which indicates a different mechanism in glyoxylate carboligases (Kaplun et al. 2008). It could also be shown that the active site 'HH-motif' which has been described for various members of the decarboxylase superfamily (Andrews and McLeish 2012) is highly specific for only a small number of decarboxylases, the pyruvate decarboxylases, indolepyruvate decarboxylases, and phenylpyruvate decarboxylases, and is not present in the majority of the enzymes. The four highly conserved glycine residues at standard positions 27, 219, 286 and 443 are all located between the C-cap of a β -strand and the N-cap of an α -helix of β - α - β supersecondary structure elements, which has been shown to be a typical pattern for α - β units (Edwards, Sternberg, and Thornton 1987). These elements presumably are relevant for the correct folding of the ThDP-dependent decarboxylases.

The assignment of standard numbers to experimentally well-characterized positions allows for an easy comparison of positions between different proteins and different organisms regarding their structural equivalence. This was demonstrated by an in-depth analysis of five different members of the decarboxylase superfamily (Supplementary Information, Table A.4 on pages 191ff.). Several positions were identified which share the same standard numbers, show similar functional influence and are structurally equivalent, but deviate in their absolute position numbers by up to 8 positions. Prediction of the functional influence of mutations in homologous sequences based on the absolute position numbers of given sequences is not straightforward, but becomes feasible using a standard numbering scheme. Thus, new sequence motifs were found by systematically analyzing the amino acid distribution at each position of all members of the ThDP-dependent

decarboxylase family. A new family-specific sequence motif was derived from the conserved region near the catalytic glutamic acid at position 51 (standard position) and the conserved alanine at position 58 (standard position). The respective motif was shown to be as sensitive and precise for the ThDP-dependent decarboxylases as the PROSITE pattern PS00187, but due to the defined E51, it cannot be used to identify glyoxylate carboligases, which have a valine at the respective position (Kaplun et al. 2008). In addition, despite the higher variability of the TH3 domain in comparison to the PYR and the PP domain, the sequence of a β -strand found in the TH3 domain (standard positions 280-286) consists of a conserved motif. In contrast to the previously mentioned motif, this region is not part of the active site but is presumably relevant for the structure or regulation of the protein. The adjacent loop region 286-304 was described as a part of the activation cascade of pyruvate decarboxylases, since this loop shows structural rearrangement upon binding of an activator at the effector binding site at standard position 221 (Kutter et al. 2009).

4.1.5 Conclusions

By introducing a robust and reliable numbering scheme for the family of ThDP-dependent decarboxylases, we provided a frame of reference for this diverse protein family. Besides being a reliable tool to identify and number residues and domain boundaries for the superfamily of ThDP-dependent decarboxylases, the presented implementation of a numbering scheme is generic and can be adapted to other protein families as well. The usefulness and reliability of the presented numbering method was demonstrated for various examples.

4.1.6 Methods

Reference alignment and position number assignment

16 representative members of the decarboxylase superfamily were selected from the ThDP-dependent Enzyme Engineering Database (Widmann, Radloff, and Pleiss 2010) by three criteria:

1. Only proteins with known crystal structure were chosen for the reference alignment. From each of the 16 homologous families that contain structure information, one member was selected.

2. Some decarboxylases show activation upon binding of a substrate molecule to an allosteric binding site which leads to conformational changes. In these cases only structures were chosen which were crystallized in complex with a bound substrate or a substrate analogue.
3. For homologues families with more than one structure entry matching these criteria, the structure with the highest resolution was selected.

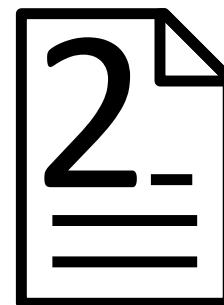
From these 16 representative proteins, a structure-guided multisequence alignment was created by STAMP (Russell and Barton 1992). This reference alignment was manually refined to align secondary structure elements and thus to reduce the number of gaps scattered in the alignment (Supplementary Information, A.4.4 on page 207). A family specific profile hidden Markov model was derived from the reference alignment by HMMER (*HMMER*, <http://hmmmer.janelia.org/> 2013). The sequence of the pyruvate decarboxylase from *S. cerevisiae* (pdb|2VK8 (Kutter et al. 2009), sp|P06169, EC: 4.1.1.1) was chosen as the reference sequence, because it is a widely applied and well characterized ThDP-dependent enzyme (Kutter et al. 2009; Pohl, Sprenger, and Müller 2004; Pohl, Lingen, and Müller 2002; Rosche et al. 2005). Standard position numbers were assigned by aligning the sequence of each member of the decarboxylase superfamily against the profile HMM and by subsequently transferring the absolute position numbers of the reference sequence to the corresponding positions of the respective decarboxylase sequence.

Web tool

An open access web application is provided to allow users to assign standard position numbers for decarboxylase sequences (www.teed.uni-stuttgart.de). After submitting a query sequence, a BLAST search against a database of members of the structural group of decarboxylases from the TEED (Widmann, Radloff, and Pleiss 2010) is performed. Only query sequences with an E-value less than 10^{-10} are accepted to guarantee for a reliable sequence alignment. Then the query sequence is aligned to the reference alignment using the profile HMM, and the absolute position numbers of the reference sequence are transferred to the query sequence. Finally, annotation information of the TEED such as catalytic residues, domain boundaries, or activator binding sites is transferred to the respective positions of the query sequence.

4.2 The modular structure of ThDP-dependent enzymes

Vogel, C; Pleiss, J. (2014) The modular structure of ThDP-dependent enzymes. *Proteins* **82** (10):2523-2537



4.2.1 Abstract

Thiamine diphosphate (ThDP)-dependent enzymes form a diverse protein family, which was classified into nine superfamilies. The cofactor ThDP is bound at the interface between two catalytic domains, the PYR and the PP domain. The nine superfamilies were assigned to five different structural architectures. Two superfamilies, the sulfopyruvate decarboxylases and α -ketoacid dehydrogenases 2, consist of separate PYR and PP domains. The oxidoreductase superfamily is of the intramonomer/PYR-PP type with an N-terminal PYR and a subsequent PP domain. The active enzymes form homodimers with the ThDP cofactor bound at the interface between a PYR and a PP domain of the same monomer. Decarboxylases are of the inter-monomer/PYR-PP type with the cofactor bound between domains from different monomers. 1-Deoxy-D-xylulose-5-phosphate synthases are of the intra-monomer/PP-PYR type. The transketolases, α -ketoglutarate dehydrogenases, and α -ketoacid dehydrogenases 1 are of the inter-monomer/PP-PYR type. For the phosphonopyruvate decarboxylases, definitive assessment of the structural architecture is not possible due to lack of structure information. By applying a structure-based domain alignment method, sequences of more than 62000 PYR and PP domains were identified and aligned. Although the sequence similarity of the catalytic domains is low between different superfamilies, seven positions were identified to be highly conserved, including the cofactor binding GDGX_{24,27}N motif, the cofactor-activating glutamic acid, and two structurally equivalent glycines in both the PYR and the PP domain. An evolutionary pathway of ThDP-dependent enzymes is proposed which explains the sequence and structure diversity of this family by three basic evolutionary events: domain recruitment, domain linkage, and structural rearrangement of catalytic domains.

4.2.2 Introduction

The family of thiamine diphosphate (ThDP)-dependent enzymes is a vast and diverse protein family found in all kingdoms of life. Because they catalyze a broad range of reactions including C-C bond cleavage and formation, ThDP-dependent enzymes are promising catalysts for biotechnological applications (Müller, Sprenger, and Pohl 2013), especially by expanding the naturally available range of enzymes by tailor-made biocatalysts with desired activity, substrate specificity, and regio- and stereoselectivity (Yep and McLeish 2009; Zhang, Dai, et al. 2003; Galman et al. 2010; Payongsri et al. 2012).

ThDP-dependent enzymes share a highly similar reaction mechanism with the cofactor ThDP as catalyst. The first step of ThDP-dependent enzyme catalyzed reactions is the formation of a ThDP ylide by protein-assisted deprotonation at its C2 atom (Andrews and McLeish 2012). Therefore, most ThDP-dependent enzymes share a highly conserved glutamic acid in the active site, which is required for cofactor activation (Shaanan and Chipman 2009; Candy, Koga, et al. 1996). In addition, the shape of the ThDP-binding pocket constrains the cofactor into its active V-shaped conformation, as observed for all ThDP-dependent enzymes (Andrews and McLeish 2012; Shaanan and Chipman 2009; Guo et al. 1998; Pletcher et al. 1977; Xiang et al. 2007; Meyer, Neumann, Koers, et al. 2012; Pang et al. 2004; Werther et al. 2010; Knoll, Müller, et al. 2006). Thus, the protein contributes to the reaction by activating the ThDP catalyst and by positioning the substrates which determines substrate specificity, regio- and stereoselectivity (Pohl, Lingen, and Müller 2002; Hailes et al. 2013).

Despite their high diversity in sequence, all ThDP-dependent enzymes share two catalytic domains, the pyrimidine- (PYR) and the pyrophosphate (PP) binding domain (Muller et al. 1993), which form the active site. All PYR and PP domains are structurally highly similar (Muller et al. 1993; Vogel, Widmann, et al. 2012; Frank, Leeper, and Luisi 2007) and share the same ThDP-binding fold (Costelloe, Ward, and Dalby 2008; Duggleby 2006), a 3-layer α - β - α sandwich with a central 6-stranded β -sheet flanked by two layers of three α -helices each. However, the ThDP-dependent enzymes differ in how the PYR and PP domains are encoded on the genes (Widmann, Radloff, and Pleiss 2010). While the two catalytic domains are on two separate genes in sulfopyruvate decarboxylases (SPDC) and members of the α -ketoacid dehydrogenase superfamily 2 (aKADH2), they are encoded in a single gene with an N-terminal PYR domain and a C-terminal PP domain in decarboxylases (DC), phosphonopyruvate decarboxylases (PPDC),

and structurally determined oxidoreductases (OR), and, in reverse order in transketolases (TK), 1-deoxy-D-xylulose-5-phosphate synthases (DXPS), the α -ketoacid dehydrogenase superfamily 1 (aKADH1), and the superfamily of the α -ketoglutarate converting enzymes (aKGDH) (Widmann, Radloff, and Pleiss 2010).

In addition, the ThDP-dependent enzymes differ by additional domains between the two catalytic domains or at the N- or the C-terminus. Based on sequence relationships, the sequential order of the two catalytic domains, and the presence or absence of additional domains, evolutionary pathways have been suggested (Costelloe, Ward, and Dalby 2008; Duggleby 2006). However, these analyses were based on only a small number of sequences and structures and did not take into account the structural architectures of the active complexes, which vary considerably. Moreover, the structural architecture might even differ between closely related superfamilies with identical sequential order of catalytic domains. The most striking example are the two homologous superfamilies TK and DXPS which consist of both an N-terminal PP domain and a subsequent PYR domain with an additional TKC-domain at the C-terminus. In both families, the active enzyme consists of a homodimer. However, despite their high overall sequence similarity, the ThDP-cofactor is bound differently. While in TKs it is bound at the interface between a PP and a PYR domain from two different monomers, in DXPSs the cofactor is bound between PP and PYR domains of the same monomer (Xiang et al. 2007). For two protein families with high sequence and structure similarity, inter- versus intramonomer localization of the cofactor might intuitively lead to the expectation of different active site interfaces in TKs and DXPSs. However, a comparison of the structures of the TK from *Saccharomyces cerevisiae* (Fiedler et al. 2002) and the DXPS from *Escherichia coli* (Xiang et al. 2007) demonstrates a close similarity of the active sites. To investigate this apparent contradiction, we systematically analyzed the sequences and structures of ThDP-dependent enzymes using the ThDP-dependent Enzyme Engineering Database (Widmann, Radloff, and Pleiss 2010) and standard numbering schemes, which have been previously applied to perform reliable multisequence alignments and to identify conserved residues in different ThDP-dependent DCs and in Cytochrome P450 monooxygenases (Vogel, Widmann, et al. 2012; Gricman, Vogel, and Pleiss 2014). The combination of a comprehensive analysis of sequence conservation and a systematic analysis of the structural architecture of different superfamilies provides a deeper insight into structural, functional, and evolutionary mechanisms of this large and diverse protein family.

4.2.3 Materials and Methods

Update of the ThDP-dependent Enzyme Engineering Database

To provide a comprehensive data source for a systematic sequence and structure analysis, the ThDP-dependent Enzyme Engineering Database (TEED) was updated based on 51 representative seed sequences chosen from 51 homologous families of the previous database version (Widmann, Radloff, and Pleiss 2010) (Supplementary Information, Table A.5 on page 210). If available, proteins with available structure information were chosen and His-tags were removed from the sequences. Subsequently, the NCBI non-redundant protein database (GenBank (Benson et al. 2011)) was searched by BLAST (Camacho et al. 2009) for sequences that are homologous to the seed sequences representing all described families of ThDP-dependent enzymes with an expectation threshold of 10^{-5} . GenBank entries with identical sequences or fragments were assigned to a single sequence entry, GenBank entries with global sequence similarities $>98\%$ were assigned to single protein entries. Homologous families were created by clustering the protein entries by usearch (Edgar 2010) with an identity cutoff of 30%. Small homologous families with less than three proteins and without structure information were deleted. Subsequently, fragments were removed by deletion of sequences shorter than 70% of the average length of the respective homologous family. Protein structures from the RCSB Protein Data Bank (PDB) (Berman et al. 2000) with a minimum sequence length of 150 amino acids and a global sequence identity of at least 80% to any sequence in the dataset of ThDP-dependent enzymes were added. In total, 77493 sequences of 52565 proteins and 240 crystal structures were parsed into the TEED. The proteins were grouped into 168 homologous families, and the 168 homologous families were classified into nine superfamilies. We assume that all members of a superfamily form enzymes with similar sequence and structure, although individual members might be single subunits such as PP or PYR domains that are part of an active heteromultimeric enzyme. Besides the eight superfamilies of the previous TEED version (Widmann, Radloff, and Pleiss 2010), the DXPS superfamily was established as a separate superfamily, which differs in structure and sequence from the TK superfamily, though its sequence similarity to the TK superfamily is higher than between other superfamilies. The updated version of the TEED is available at www.TEED.uni-stuttgart.de.

Structure alignment

To identify the borders of the PYR and the PP domain, 144 crystal structures of representative ThDP-dependent enzymes were compared to the structures of the pyruvate decarboxylase from *S. cerevisiae* (pdb|2VK8) in PyMOL (*The PyMOL Molecular Graphics System, Version 1.3*, Schrödinger, LLC, <http://pymol.org> 2013) (Supplementary Information, Table A.6 on page 213). The PYR and PP domains were defined between standard positions 6 and 168 and between 367 and 540, respectively, according to the standard numbering scheme for ThDP-dependent DCs (Vogel, Widmann, et al. 2012). Subsequently, the respective domains were aligned by STAMP (Russell and Barton 1992). By mapping the overall structure of all available crystal structures of ThDP-dependent enzymes on the respective PP domains in PyMOL, an alignment of all available structures was generated and manually fine-tuned.

Domain-based standard numbering scheme and multisequence alignments

Based on the standard numbering scheme for ThDP-dependent DCs (Vogel, Widmann, et al. 2012), a domain-based numbering scheme for all ThDP-dependent enzymes was developed in order to allow for the automatic identification of PYR and PP domains in all sequences. To derive profile hidden Markov models (HMM) for PYR and PP domains that are sufficiently variable to cover the wide sequence variation of ThDP-dependent enzymes, initial structure alignments of 50 and 48 representative PYR and PP domains were generated by STAMP (Russell and Barton 1992). Subsequently, sequences from homologous families without structure information were added. Finally, profile HMMs for the two domains were derived from the multisequence alignments using hmmbuild from the HMMER package (Johnson, Eddy, and Portugaly 2010). In order to locate the PYR and the PP domain, hmmsearch was applied to each sequence in the TEED to identify the matching sequence part with the highest score, an E-value $< 10^{-4}$ and a minimum length of 100 amino acids for each of the two domain-specific profiles. If the two hits for the two domains overlapped on the query sequence, two cases were distinguished to obtain an unambiguous assignment of PYR or PP domains:

1. The longer hit was kept and the smaller hit was discarded, if they overlapped by less than 60% of the longer hit.
2. Both hits were discarded if they overlapped by more than 60% of the longer hit.

This approach resulted in 63108 and 62035 sequences of PYR and PP domains, respectively. For each sequence, standard numbers were assigned to each position in the two identified domains by transfer of the corresponding position number from the reference sequence, the pyruvate decarboxylase from *S. cerevisiae* (Vogel, Widmann, et al. 2012). Subsequently, the PYR and PP domains were separately aligned by arranging all positions with the same standard numbers in equivalent alignment columns. Thus, the domain-based standard numbering scheme allowed for a reliable multisequence alignment which is computationally efficient due to its linear scaling with the number of sequences.

Analysis of sequence relationships

The sequence similarities between the PYR and PP domains of a majority of ThDP-dependent enzymes were calculated from separate multisequence alignments of the two domains which were generated using the domain-based numbering scheme. Each sequence pair was analyzed for the percentage of similar columns. A corresponding position in a sequence pair was defined as 'similar' if its BLOSUM 62 substitution score was ≥ 0 . The similarity of a pair of sequences was calculated as the ratio between the number of similar positions and the total number of positions (excluding gap-containing positions). Average sequence similarities were calculated for all pairs between two homologous families, and a similarity matrix for PYR as well as for PP domains of all homologous families with identified catalytic domains was derived. Using *Cytoscape* (Smoot et al. 2011), networks with nodes representing the homologous families and edges representing average sequence similarities above 50% were generated. Additionally, in order to analyze the relative conservation of PYR and PP domains in ThDP-dependent enzymes with fused PYR and PP domains, the similarities of those domains between all pairs of 49347 sequences of fusion proteins were calculated. Thus, for each combination of sequence similarities of PYR and PP domains (intervals of 1%), the number of pairs was counted and visualized as a heatmap using *gnuplot* 4.4.

Secondary structure naming scheme

A naming scheme for secondary structure elements was proposed based on the ThDP-binding fold (Duggleby 2006). α -Helices were named by letters (α A, α B, ...), β -strands by numbers (β 1, β 2, ...). To differentiate between secondary structure elements from the two catalytic domains,

the prefix 'PYR' or 'PP' was added.

4.2.4 Results

Modular structure of ThDP-dependent enzymes

The Thiamine diphosphate-dependent Enzyme Engineering Database (TEED, www.TEED.uni-stuttgart.de) was updated and resulted in 77493 sequences of 52565 proteins. The proteins were assigned to 168 homologous families and nine superfamilies (Table 4.2 on the next page). Seven superfamilies contained proteins with structure information, for only two superfamilies (SPDC and PPDC) structure information is lacking. To systematically analyze the architecture of ThDP-dependent enzymes, the 240 structures included in the TEED were compared.

In addition, the catalytic PYR and PP domains were automatically identified in a majority of sequences in the TEED using domain specific profile HMMs. In two superfamilies, the majority of enzymes consist of PYR and PP domains coded by separate genes; in seven superfamilies the majority of enzymes have the PYR and the PP domain encoded on a single gene, however in different sequential order (PYR-PP or PP-PYR). Separated PYR and PP domains were identified for a majority of sequences from the superfamilies of SPDCs and the α -ketoacid dehydrogenase superfamily 2 (aKADH2). Both catalytic domains fused on one gene were found in a majority of sequences from the superfamilies of DC, TK, DXPS, OR, PPDC, the aKGDH, and the aKADH1. In the ORs which have the catalytic domains fused on one chain, as well as in DCs and PPDCs, the sequential order of the catalytic domains is PYR-PP, while it is reversed (PP-PYR) in TKs, DXPSs, aKADH1s, and aKGDHs.

Additional information about the architecture of the active enzyme complex was derived from crystal structures of ThDP-dependent enzymes. In two superfamilies with available structure information, the OR and DXPS, the cofactor ThDP is bound at the interface between a PYR and PP domain of the same protein monomer ('intra-monomer'), whereas in the superfamilies of DC, TK, aKADH1, and aKGDH, the cofactor is bound between a PYR and a PP domain from different monomers ('intermonomer'). Thus, the seven superfamilies of ThDPdependent enzymes with available structure information can be assigned to five different structural types (Figure 4.3 on page 100): separate PYR and PP domains [Figure 4.3(A)], two intramonomer types with sequential order PYR-PP [Figure 4.3(B)] or PP-PYR [Figure 4.3(D)], and two inter-monomer

types with sequential order PYR-PP [Figure 4.3(C)] or PP-PYR [Figure 4.3(E)]. Despite their seemingly different quaternary structure, however, in all ThDP-dependent enzymes with structure information the cofactor ThDP is bound in similar binding pockets, and the architecture of the active site is highly conserved. A superposition of representative structures of seven superfamilies with structure information revealed that the four PYR and PP domains forming the active protein are arranged identically with two molecules of ThDP bound between the two PYR/PP pairs. The major difference between the five structural types is their 'wiring' - how the PYR and PP domains are connected by loops and additional domains (Figure 4.4 on page 101).

Table 4.2: Superfamilies of ThDP-dependent enzymes and their functional annotation in the NCBI GenBank. For each superfamily, multiple homologous families were established in the TEED based on sequence similarity.

Superfamily	Functional annotation
DC	pyruvate oxidases, pyruvate dehydrogenases (cytochrome), indolepyruvate decarboxylases, phenylpyruvate decarboxylases, pyruvate decarboxylases, branched-chain α -ketoacid decarboxylases, 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthases, isochorismate synthases, PHYLLLO enzymes, benzoylformate decarboxylases, benzaldehyde lyases, oxalyl-CoA decarboxylases, 2-hydroxyphytanoyl-CoA lyases, acetohydroxyacid synthases, acetolactate synthases, sulfoacetaldehyde acetyltransferases, carboxyethylarginine synthases, glyoxylate carbonylases, cyclohexane-1,2-dione hydrolase, YerE-like enzymes, PigD-like enzymes, 3D-(3,5/4)-trihydroxycyclohexane-1,2-dione hydrolases
TK	transketolases, phosphoketolases, dihydroxyacetone synthases
DXPS	1-deoxy-D-xylulose-5-phosphate synthases
OR	pyruvate:ferredoxin oxidoreductases, 2-ketoglutarate:ferredoxin oxidoreductases, indolepyruvate:ferredoxin oxidoreductases, 2-ketoisovalerate:ferredoxin oxidoreductases
aKADH1	E1 components of pyruvate dehydrogenase
aKADH2	branched-chain α -keto acid dehydrogenase E1 components, 2-oxoisovalerate dehydrogenases, acetoin:2,6-dichlorophenolindophenol oxidoreductases
SPDC	sulfoxyruvate decarboxylases
PPDC	phosphonopyruvate decarboxylases
aKGDH	E1 components of 2-oxoglutarate dehydrogenases, 2-oxoglutarate decarboxylases

Separated PYR and PP domains

In all SPDCs and in the majority of aKADH2s, the PP and PYR domains are coded in different genes [Figure 1(A)]. 49% and 44% of the aKADH2 sequences were found to be PP and PYR domains, respectively, 5% encoded both domains in the order PP-PYR (Table 4.3 on page 102).

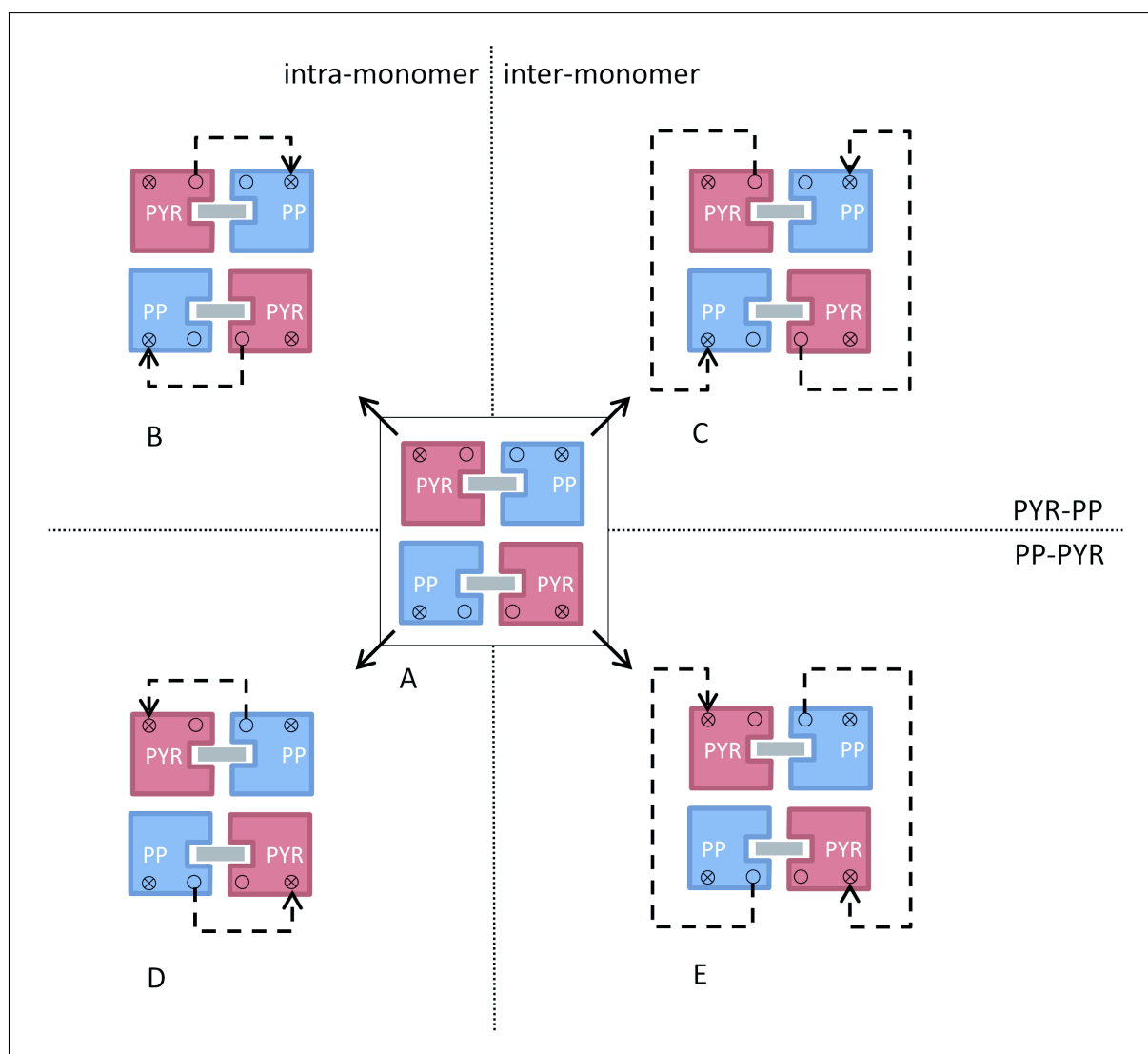


Figure 4.3: The five different basic layouts found in the known structures of ThDP-dependent enzymes. All structurally determined ThDP-dependent enzymes contain PYR and PP domains forming the active sites. Besides no coupling of those domains on the sequence level (A), both domains can be coupled in four different types depending on the order of the PYR and PP domain on the sequence and the linkage in the structure. Structural types with the cofactor bound between the PYR and PP domains fused on one protein chain are called 'intra-monomer' (B,D), whereas types with the cofactor bound between a PYR and a PP domain of different protein chains are called 'inter-monomer' (C,E). In the structures of ThDP-dependent enzymes, type A was exclusively found in aKADH2 but sequence analysis further revealed 38% of all TK sequences and ~60% of all OR sequences to belong to this type. Type B was found for the majority of ORs encompassing PFOR and IOR and type C was found to be characteristic for the DC superfamily. The structures of DXPSs belong to type D, whereas TKs, aKADH1s, and aKGDHs belong to type E.

Due to sequence similarity between aKADH2s and TKs, those proteins most probably are of the inter-monomer/PP-PYR type. Interestingly, the aKADH2 superfamily is not the only family that contains members with separated and members with fused PP and PYR domains. Minorities with separated domains were also found in the TK, PPDC, and OR superfamilies.

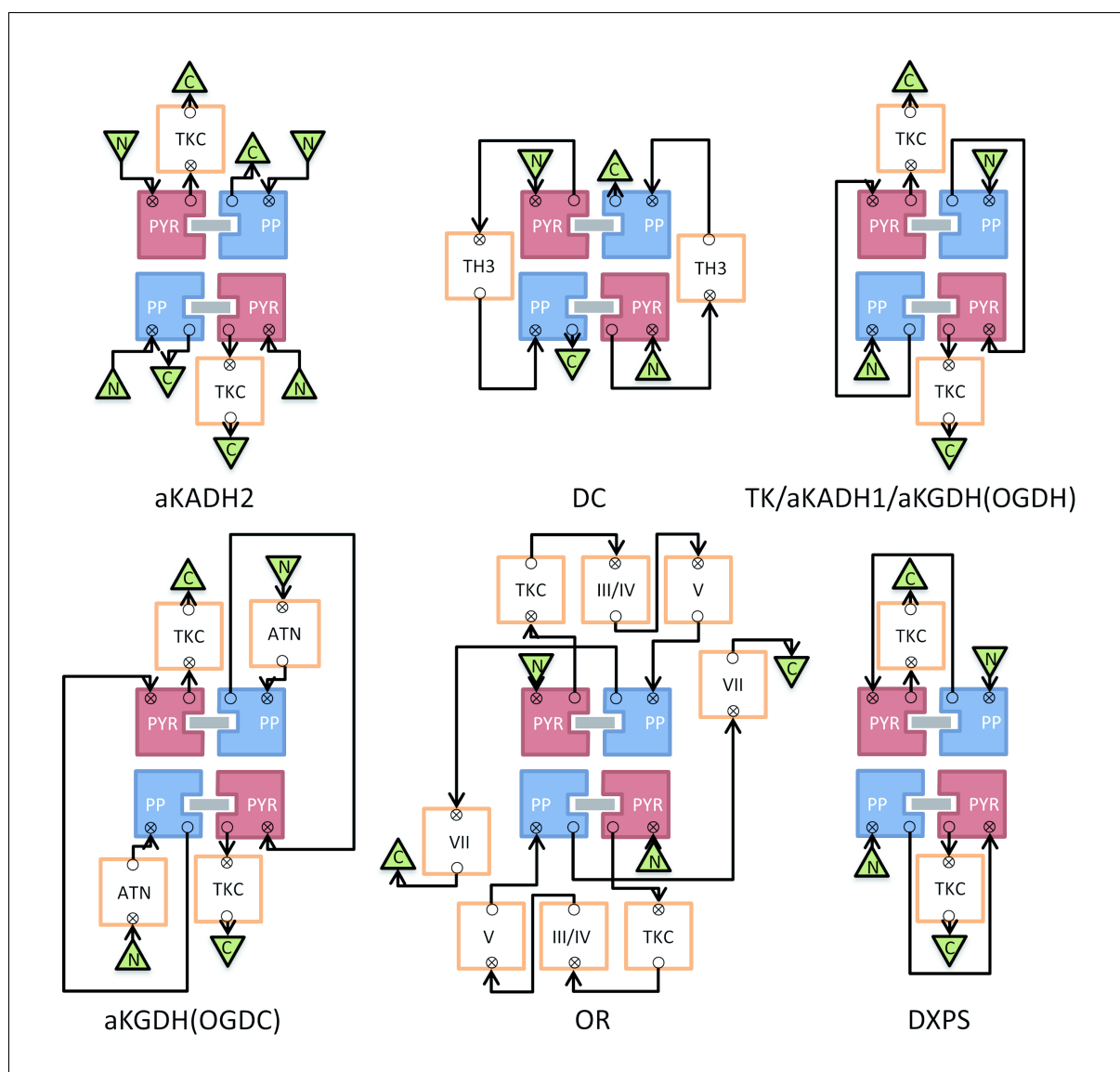


Figure 4.4: Structural architectures found in ThDP-dependent enzymes with available structure information. All ThDP-dependent enzymes with available structure information share a conserved structural orientation of the catalytically relevant PYR and PP domains, while six different family-specific domain arrangements can be observed. DXPSs, ORs, aKADH2s, and DCs were classified as distinct superfamilies based on global sequence similarity between their homologous sequences and their specific domain arrangement. TKs and aKADH1s share their domain arrangement with OGDHs but were separated as superfamilies due to clear separation on sequence level. OGDHs and OGDCs were grouped to the aKGDH superfamily based on global sequence similarity, although the OGDCs have an extended domain layout by implementation of an ATN domain.

Intra-monomer/PYR-PP type

In the OR superfamily, crystal structures are exclusively available for members of the subfamily of pyruvate: ferredoxin oxidoreductases (PFOR). Those structures show the most complex domain architecture found in ThDP-dependent enzymes combining the PYR and PP domains with four additional domains. The N-terminal PYR domain is linked to the subsequent PP domain

Table 4.3: Results of the automatic domain identification in all ThDP-dependent enzymes. For each superfamily, the distribution of detected domains and the average length of the catalytic domains is shown.

SFAM	percentage ¹				average domain length ²	
	PYR / PP	PYR	PP	no hit	PYR	PP
DC	96	2	2	1	154 +/- 9	161 +/- 17
TK	62	19	19	0	150 +/- 16	193 +/- 15
DXPS	99	0	0	0	132 +/- 11	202 +/- 27
OR	32	36	27	5	156 +/- 7	190 +/- 48
aKADH1	98	0	0	2	185 +/- 9	264 +/- 10
aKADH2	5	44	49	1	156 +/- 12	183 +/- 17
SPDC	0	33	61	6	139 +/- 15	126 +/- 14
PPDC	70	21	6	3	142 +/- 19	141 +/- 17
aKGDH	99	0	1	0	190 +/- 8	216 +/- 11

PYR/PP: both domains were detected in x% of the sequences from this superfamily.

PYR: in x% of the sequences from this superfamily, only the PYR domain was identified.

PP: in x% of the sequences from this superfamily, only the PP domain was identified.

no hit: in x% of the sequences from this superfamily, no ThDP-binding domain was identified

¹ Deviation of the sum of all percentages from 100 is due to rounding.

² Average length of the PYR and PP domains and standard deviation.

via three domains; the TKC domain which is similar to the C-terminal domain of transketolases (TKC), the domain III/IV described previously to consist of two individual domains III and IV (Chabrière et al. 1999), and domain V. In addition, domain VII is C-terminally linked to the PP domain. The cofactor ThDP is bound at the interface between a PP and a PYR domain of the same monomer (intramonomer) [Figure 4.3(B) on page 100]. An PYR-PP type with additional domains was also observed for members of the indolepyruvate:ferredoxin oxidoreductases (IOR), which together with the PFOR constitute one half of all OR sequences. As already mentioned, the OR superfamily additionally contains enzymes with their PYR and PP domains coded by separate genes (making up 36% and 27% of all OR entries, respectively). Besides, the mainly PYR-PP type PFOR and IOR, the OR superfamily encompasses the solely $\alpha_2\beta_2$ -heterotetrameric 2-ketoisovalerate:ferredoxin oxidoreductases (VOR) and 2-ketoglutarate oxidoreductases (KGOR). As a specific feature of ORs, some of the family members contain their domain VII (γ domain) on an additional chain, which is represented by the number of 5% of OR sequences without any of the catalytic domains PYR and PP.

Inter-monomer/PYR-PP type

In the DC superfamily, the sequential order is similar to the fused ORs (PYR-PP), while the cofactor is bound at the interface between two dimers (inter-monomer) [Figure 4.3(C)]. The N-terminal PYR domain is linked to the C-terminal PP domain by an additional TH3 domain in all DCs.

Intra-monomer/PP-PYR type

The two members of the DXPS superfamily with available structure information are of the intra-monomer/PP-PYR type [Figure 4.3(D)] and have their N-terminal PP domain linked to the subsequent PYR domain by a short loop. The PYR domain is C-terminally followed by a TKC domain.

Inter-monomer/PP-PYR type

Three superfamilies are of the inter-monomer/PP-PYR type [Figure 4.3(E)]; the superfamilies TK, aKADH1, and aKGDH, the latter encompassing the E1 components of 2-oxoglutarate dehydrogenases (OGDH) and 2-oxoglutarate decarboxylases (OGDC). In all three superfamilies, the N-terminal PP domain is linked to the PYR domain via a long loop of 120 residues on the average. The PYR domain is followed by a C-terminal TKC domain. The OGDC family contains an additional N-terminal acyltransferase-like domain (ATN). The two superfamilies TK and aKADH1 have high sequence similarity to the intra-monomer type DXPS. The major difference between TKs and aKADH1s, which are of the inter-monomer/PP-PYR type, and the DXPSs, which are of the intra-monomer/PP-PYR type, is the length of the linker connecting the PP and the PYR domain (average length of 120 and 58 residues, respectively). Thus, small sequence changes such as the loop length might be sufficient to switch between the inter- and the intramonomer type. In all superfamilies, the fraction of sequences without any identified catalytic domain was below 6%. In those cases, the profiles identified sequence fragments not matching the criteria of a minimum length of 100 amino acids and of an unambiguous assignment of a PYR or a PP domain. In addition, a small number of sequences did not contain a catalytic domain but consisted of additional domains such as the γ domain as described for ORs.

Superfamily-specific structural variations in PYR and PP domains

All PYR and PP domains show overall structural similarity and belong to the ThDP-binding fold. The ThDP-binding fold consists of a central parallel 6-stranded β -sheet with the β -strands connected by α -helices, thus forming a three-layered α - β - α sandwich structure (Figure 4.5 on the next page). The active pocket-lining residues of the PYR and the PP domain are located in three loop regions between β -strand β 1 and α -helix α B, between β 2 and α C, between β 3 and α D, and in the PP domain additionally in the region between β 4 and α F including α E. The structures of the PYR domains of all superfamilies are highly similar and differ in only three regions.

Region 1: The loop between PYR- α B and PYR- β 2 is long in structures of aKGDHs, aKADH1s, ORs, and the non-human TKs and is short in structures of DCs, DXPSs, aKADH2s, and the human TKs. The phosphoketolases (treated as a subfamily of the TK superfamily so far) deviate in the orientation of this loop compared to the remaining TK structures.

Region 2: Additionally, a small antiparallel β -sheet consisting of two short β -strands is orthogonally inserted after PYR- α B in aKGDHs followed by a partially helical loop, while in all other superfamilies of ThDP-dependent enzymes, PYR- α B is connected to PYR- β 2 by a smaller loop.

Region 3: In the DC superfamily, an α -helix PYR- α E is present, while in all other ThDP-dependent enzymes, the respective region is a loop.

The structures of PP domains differ in four regions, the two variable regions 1 and 3 of the PYR domains and two additional regions 4 and 5. *Region 1:* In TKs, DXPSs, aKGDHs, and in both homologous families of aKADHs, PP- α B is followed by an insertion consisting of an helix and a long coil, while in DCs and ORs, PP- α B and PP- β 2 are linked by a short loop. In aKGDHs, this helix-loop insertion is followed by an additional β -strand antiparallel to PP- β 2. In combination with an N-terminal extension in length of PP- β 2, the central β -sheet is larger than in the common ThDP-binding fold.

Region 3: As observed for the PYR domains, the existence of PP- α E is specific for the DC superfamily. In TKs, the loops between PP- β 4 and PP- α F are shortened as compared to the DC superfamily and PP- α E and the subsequent loop are replaced by a short antiparallel β -sheet. In DXPSs, aKADHs, aKGDHs, and ORs, PP- α E is replaced by a loop.

Region 4: In ORs, helix PP- α C is drastically elongated and forms a 4-helix bundle with three additional helices. This additional structural element emerges out of the protein.

Region 5: The aKADH1s additionally obtain a large insertion between PP- β 5 and PP- α G consisting of six short α -helices connected by short loops. This additional element emerges out of the protein structure.

The observed structural variation of the PYR and PP domains in different superfamilies are reflected on sequence level by the length of the respective domains (Table 4.3 on page 102). In accordance with the smaller deviation between structures of the PYR domains as compared to the PP domains, less variation in the domain length was observed for the sequences of the PYR domains.

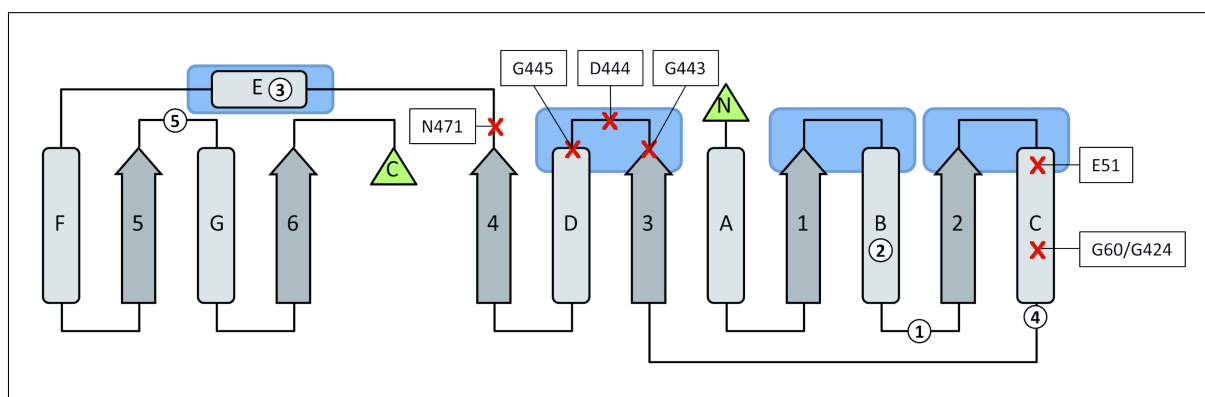


Figure 4.5: Naming scheme for the ThDP-binding fold conserved within the PYR and PP domains of ThDP-dependent enzymes. The fold representation was derived from Ref. (Duggleby 2006). The numbers 1-5 define the deviating secondary structure composition between members of different superfamilies of ThDP-dependent enzymes. The blue highlighted regions are part of the active site. α -Helix E only contributes to the active site in PP domains. Red crosses mark the positions of the seven highly conserved ($\geq 80\%$) residues in all proteins that form part of ThDP-dependent enzymes. E51 and G60 are located within the PYR domain, G424, G443, D444, G445, and N471 are located within the PP domain.

Sequence similarity of PYR and PP domains

Two multisequence alignments of 63108 PYR domains and 62035 PP domains from 110 and 132 homologous families, respectively, were performed. The average pairwise sequence similarities between homologous families were calculated separately for the PYR and PP domains. From the distance matrices, similarity networks of the PYR and the PP domains were constructed. The nodes represent homologous families, the edges an average sequence similarity of more than 50% (Figure 4.6 on page 108). By applying a force-directed layout, the nodes were grouped into clusters which are highly linked and separated from less similar sequences. The sequence similarity networks for the PYR and PP domains both show clustering of homologous families belonging to distinct superfamilies.

PYR domains: The PYR domains [Figure 4.6(A)] fall in five clearly separated clusters mainly represented by the six superfamilies DC, SPDC/PPDC, OR, aKADH2, and TK. The smaller superfamilies DXPS, aKADH1, and aKGDH are associated to other clusters: DXPSs connect to TKs and aKADH2s, aKADH1s to TKs, and aKGDHs to aKADH2s. Notably, this clustering does not correlate with the variations in regions 1 and 2 of the PYR domains, but may be mediated by the variations in region 3 (Supplementary Information, Table A.7 on page 215). Inside the clusters, the nodes are highly interlinked which indicates a high pairwise similarity of all members of each cluster. Moreover, between the clusters there are only few linkages, indicating a small number of homologous families that have high sequence similarity though belonging to different clusters. No linkage was observed for the homologous family containing the PigD enzymes (homologous family 25), which belongs to the DC superfamily by global sequence similarity. Two of three families of phosphoketolases (PK) (homologous families 28 and 72), which are globally similar to TK enzymes, are neither connected to each other nor to the TK superfamily cluster whereas the third PK family (homologous family 104) is part of the cluster of TK.

PP domains: For the PP domains [Figure 4.6(B)], clustering resulted in a more interlinked network of superfamily-specific subclusters. With the exception of the superfamilies PPDC and SPDC, the superfamilies form highly connected subclusters. The PPDCs and SPDCs are highly linked to members of the DC superfamily, but additionally connected to TK and OR families. The most prominent relationship was observed for the DC, TK, and aKADH2 superfamilies, which are connected via many, homologous families. Less connections between different superfamilies and therefore more distinct affiliation to specific clusters was found for the OR, aKGDH, aKADH1, and DXPS superfamilies. The ORs connect to DCs and to TKs, aKGDHs to aKADH2s, aKADH1s to TKs and DXPSs to TKs and aKADH2s. As for the PYR domains, the PigD family and two of three PK families (homologous families 72 and 104) were not connected to any other family. In contrast to the PYR domains, the PP domains of homologous family 28 (phosphoketolases) are part of the TK cluster.

Furthermore, in order to evaluate the relative conservation of both catalytic domains, their similarity in the PYR-PP and PP-PYR type fusion proteins was analyzed (Figure 4.7 on page 109). The majority of sequence pairs showed similarities of 40% and 46% for their PYR and PP domains, respectively, but at higher sequence similarity, the PYR domains revealed to be more conserved than the PP domains. Moreover, seven positions were highly conserved in more than

80% of the domain sequences. Two conserved positions are located in the PYR domain: the catalytic glutamic acid at standard position 51 (numbering according to the previously introduced standard numbering scheme for ThDP-dependent DCs (Vogel, Widmann, et al. 2012)) and a glycine at standard position 60. Both positions are at the N-cap and in the center, respectively, of helix PYR- α C and are conserved in more than 98% and 82%, respectively, of all sequences. In addition, five highly conserved positions are located on three different secondary structure elements in the PP domain. The first three residues of the cofactor-binding GDGX_{24,27}N motif (Hawkins, Borges, and Perham 1989) (standard positions 443, 444, and 445) form the bond between PP- β 3 and PP- α D and were found to be conserved in 96%, 96%, and 80%, respectively, of all sequences. The asparagine of this motif at standard position 471 is conserved in 91% of all identified sequences and is located in the loop between PP- β 4 and PP- α E. The fifth conserved residue is a glycine at standard position 424 in PP- α C (88% of all sequences). A structural superimposition of the PYR and PP domains of the pyruvate decarboxylase from *S. cerevisiae* (ScPDC; pdb|2VK8) revealed the two conserved glycines at standard positions 60 and 424 to be structurally equivalent.

4.2.5 Discussion

Conservation of PYR and PP domains and functional implications

As compared to the previously published version of the TEED (Widmann, Radloff, and Pleiss 2010), the update resulted in a sixfold increase in the numbers of sequences and proteins. This increase is both due to a more comprehensive sequence search and to a drastic increase of available sequence data from genome projects that identified ThDP-dependent enzymes in all kingdoms of life. ThDP-catalyzed reactions were assumed to play an important role since the earliest stages of life (Frank, Leeper, and Luisi 2007) and are of vital importance for all contemporary organisms. Although the ubiquitous protein family of ThDP-dependent enzymes is highly diverse and encompasses nearly 80000 known sequences, seven positions were found to be highly conserved. A functional relevance was described for five positions, the cofactor-activating glutamic acid at standard position 51 (Shaanan and Chipman 2009; Candy, Koga, et al. 1996; Lee, Lee, et al. 2013), and the cofactor-binding GDGX_{24,27}N motif (standard positions 443, 444, 445, and 471) (Hawkins, Borges, and Perham 1989). In addition, the conservation analysis revealed two glycines with conservation of 82% and 88%, respectively, at two structurally equivalent

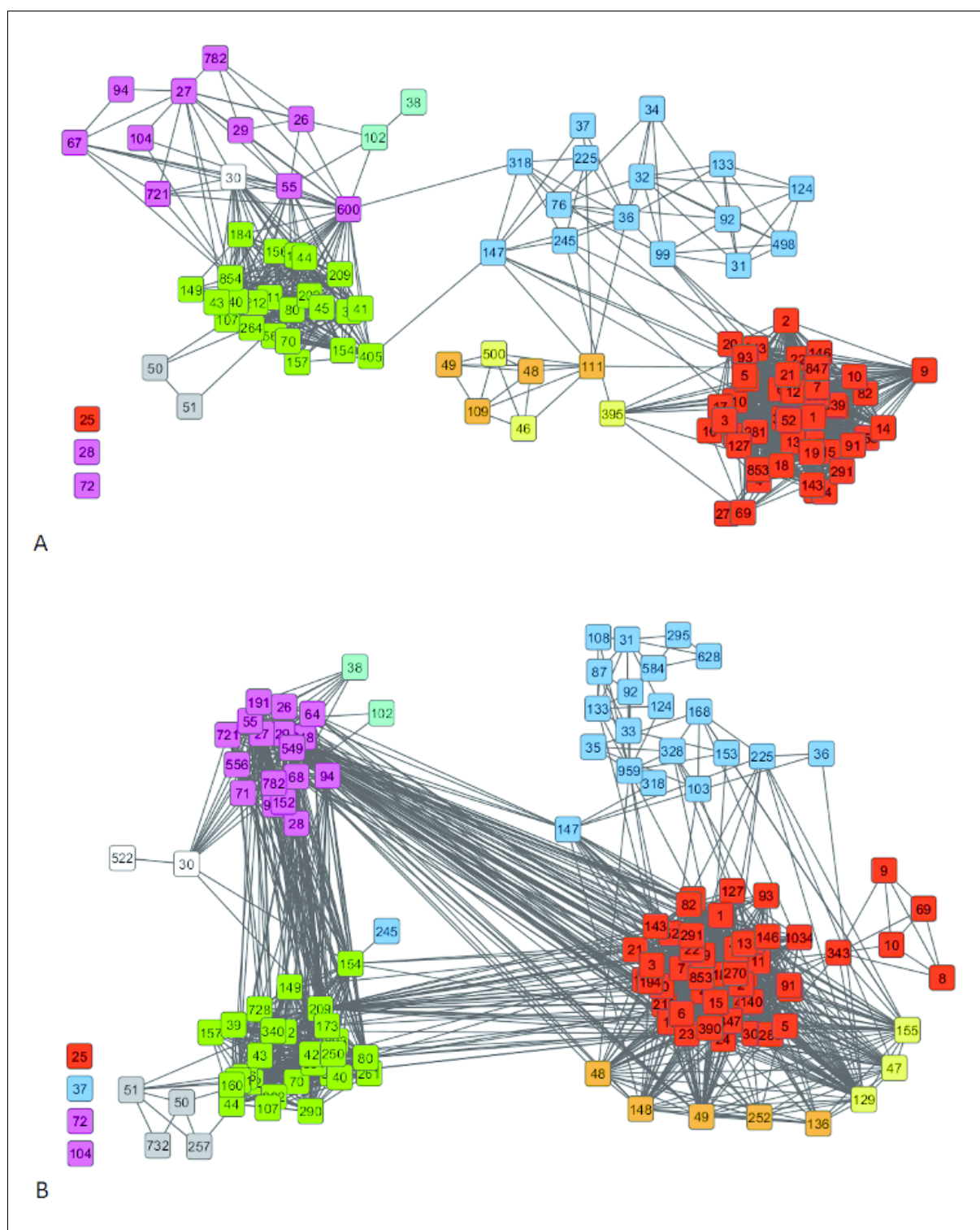


Figure 4.6: Network representation of the average sequence similarities between the PYR domains (A) and the PP domains (B) of 110 and 132 homologous families, respectively, of ThDP-dependent enzymes. The nodes represent the homologous families of the TEED connected by edges showing average sequence similarities above 50%. The coloring represents the membership to the different superfamilies DC (red), TK (pink), DXPS (white), OR (blue), SPDC (yellow), PPDC (orange), aKADH1 (cyan), aKADH2 (green), and aKGDH (gray).

standard positions 60 and 424 in the PYR and PP domain, respectively. Both residues are not part of the active site and are positioned in the center of helix αC , thus most likely they play a

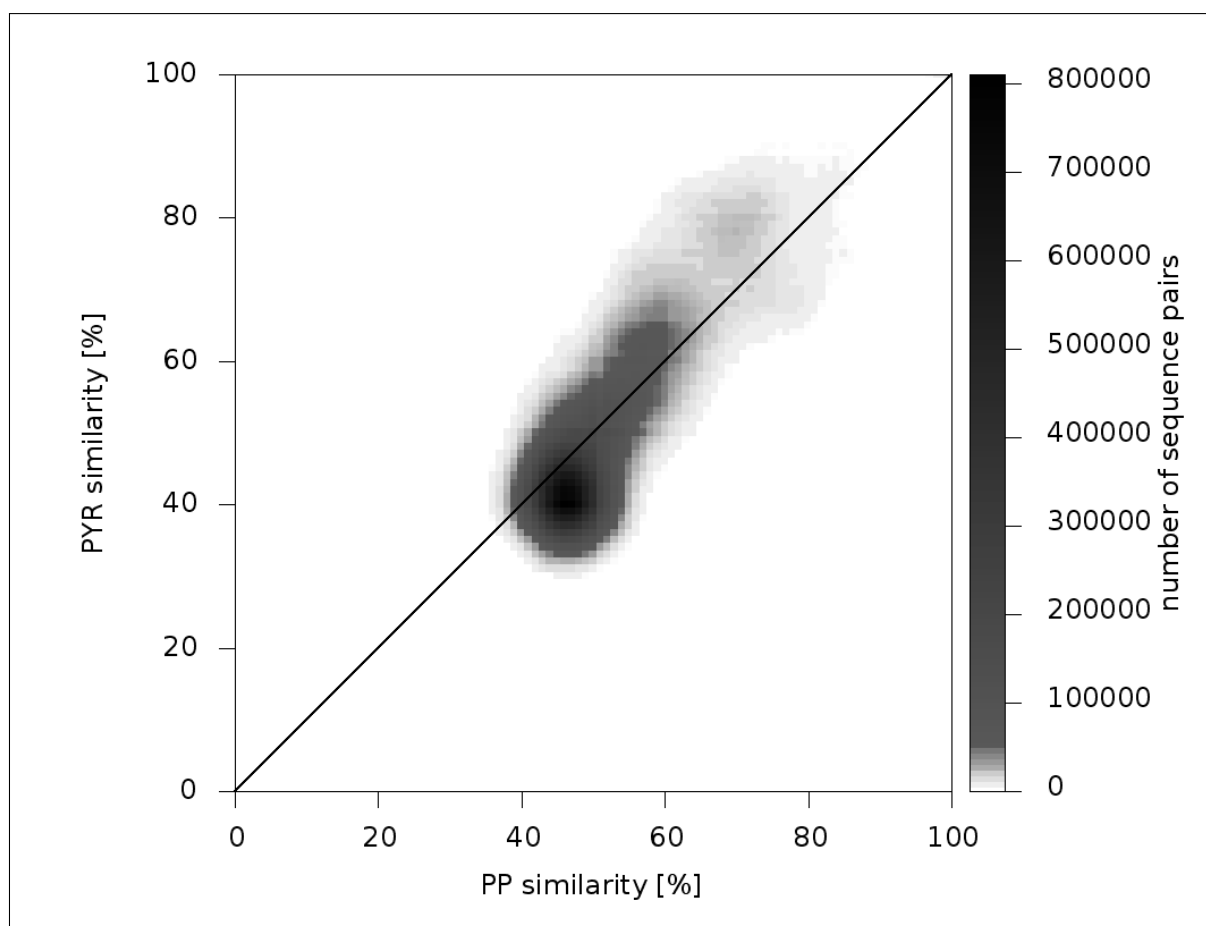


Figure 4.7: Relative average sequence similarity of the PYR and PP domains between the PYR and PP domains of 49347 PYR-PP or PP-PYR type fusion proteins. The color scale from white to black reflects the number of sequence pairs with defined similarities of their PYR and PP domains.

crucial but yet unknown role in folding or stability of the ThDP-binding fold.

Apart from the seven highly conserved positions and the structural similarity of the PYR and the PP domains, variations of the size (Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008) and the physicochemical properties of the active site pocket (Kaplun et al. 2008) were observed, which corresponds to the vast diversity of substrate and product range (Frank, Leeper, and Luisi 2007) as well as stereo- and regioselectivity (Hailes et al. 2013). The major difference in the active site pocket is the insertion of helix PP- α E in DCs, which is lacking in the other ThDP-dependent enzymes. Because stereoselectivity-mediating positions in DCs, the *S*-pocket (Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008), are located on helix PP- α E, this helix is a pivotal element that determines stereoselectivity in the DC superfamily. An exchange of amino acids located on PP- α E resulted in a change of stereoselectivity in the carbonylation of non-natural substrates with variants of the pyruvate decarboxylase from *Acetobacter pasteurianus* (Rother et al. 2011), the benzoylformate decarboxylase from *Pseudomonas putida* (Gocke, Walter, et al.

2008), and the MenD enzymes from *Escherichia coli* (Westphal, Hahn, et al. 2013; Westphal, Waltzer, et al. 2013) and *Bacillus subtilis* (Westphal, Jansen, et al. 2014). The lack of PP- α E in other superfamilies or the replacement by a short antiparallel β -sheet in TKs might explain the functional differences such as stereoselectivity among DCs, TKs, and other superfamilies (Hailes et al. 2013). For the four variable regions outside the active site pocket, structural or functional implications are still under discussion. For the flexible insertion of an additional β -strand and an α -helix after PP- α B in aKGDHs, participation in binding an AMP molecule was described, but the structural or functional role of nucleotide binding remained obscure (Frank, Price, et al. 2007). The short, antiparallel β -sheet after PYR- α B in aKGDHs contacts an helix-turn-helix motif, which is part of the linker between the PP and PYR domains of the respective enzymes. This insertion most probably stabilizes the structure of the domain-connecting linker. No structural role was found for the additional helix after PP- α B in TKs, DXPSs, aKADHs, and aKGDHs, which is located at a position where the TH3 domain is located in DCs or the domain III/IV in ORs. The extensive insertion by 70 residues after PP- β 5 of aKADH1 proteins is located at the protein surface near the N-terminus of the protein and thus might influence substrate binding, allosteric activation or inhibition, or the assembly in multienzyme complexes (Duggleby 2006). Thus, deviations from the general ThDP-binding fold imply mostly a structural rather than a functional role.

Wiring of the PYR and PP domains

For all ThDP-dependent enzymes with known structure, an almost identical core structure consisting of two PYR and two PP domains forming two active sites was found (Muller et al. 1993; Frank, Leeper, and Luisi 2007). Nevertheless, the global sequence similarity is generally low between different superfamilies, and the length of the sequences and the order of the PYR and PP domains varies (Widmann, Radloff, and Pleiss 2010). Based on the sequential order of the catalytic domains, ThDP-dependent enzymes can be distinguished into PYR-PP types, PP-PYR types, and enzymes with separated domains. Moreover, the active site can be composed by PYR and PP domains from the same or two separate monomers. In order to classify ThDP-dependent enzymes by their architecture, we introduced the concept of intra- and inter-monomer types. Intuitively, the observation of homologous enzymes with their active site within the monomer (intra-monomer) and enzymes with their active site at the dimer interface (inter-monomer) seems to be incompatible with the assumption of a highly similar core structure shared by all

ThDP-dependent enzymes. This is illustrated by the TK and DXPS superfamilies, which have a high overall sequence similarity, but a different architecture. Enzymes of both families form active dimers of the PP-PYR type (Xiang et al. 2007; Fiedler et al. 2002), and the active sites of TKs and DXPSs are highly similar. However, in TKs the active site is at the interface between two monomers (inter-monomer), while in DXPSs it is within each monomer (intra-monomer). This apparent contradiction is the consequence of a structural rearrangement of the PYR and the TKC domain in DXPSs as compared to TKs, while keeping the sequential order of the domains unchanged. This structural rearrangement can best be explained by a 'rewiring' of the three domains (Figure 4.3 on page 100).

Structure of sulfopyruvate decarboxylases and phosphonopyruvate decarboxylases

The SPDC and PPDC superfamilies have a high sequence similarity. Most SPDC genes encode separate PYR and PP domains, but a minority of SPDCs such as the SPDC from *Methanosarcina acetivorans* is of the PYR-PP type and was shown to have SPDC activity (Graham et al. 2009). Based on sequence similarity, this SPDC was assigned to the PPDC superfamily of the TEED, reflecting the high sequence similarity between both families. In contrast to SPDCs, most of the PPDC genes are of the PYR-PP type. SPDCs and PPDCs are the only two superfamilies without experimentally derived structure information, and information on quaternary structure and thus on the architecture of the active protein is restricted to results from gel filtration chromatography. For the SPDC enzymes from *Methanococcus jannaschii* and *Roseovarius nubinhibens*, the two separate PYR and PP domains form a $\alpha_6\beta_6$ -heterododecameric complex (Graupner, Xu, and White 2000; Denger et al. 2009). For the PYR-PP type PPDC from *Bacteroides fragilis*, a homotrimeric form was proposed (Zhang, Dai, et al. 2003), whereas a homodimeric form was suggested for the PYR-PP type PPDC from *Streptomyces viridochromogenes* (Johnen and Sprenger 2009). From these results, four different structural types are possible: an inter-monomer dimer, an intra-monomer dimer, an inter-monomer trimer, and an intra-monomer trimer.

Homodimeric PPDC: PPDCs have a short linker between the PYR and the PP domains. Therefore, in analogy to TKs and DXPSs, the intra-monomer type is more probable than the inter-monomer/PYR-PP type. The putative structural architecture of a homodimeric, inter-monomer/PYR-PP type PPDC would then be similar to the architecture of the DC superfamily, where the longer distance between the catalytic domains is bridged by an additional TH3 domain.

A long linker, but in reverse direction connecting the N-terminal PP to the subsequent PYR domain, was found in the inter-monomer/PP-PYR type of TKs, aKADH1s, and aKGDHs. The analysis of the intermonomer type TK and aKADH1 sequences revealed an average linker length of 120 residues, as compared to an average of 58 and 54 residues in the intra-monomer DXPS and PPDC sequences, respectively.

Homotrimeric PPDC: The architecture of a putative homotrimeric PPDC from *B. fragilis* remains unclear, since no structurally corresponding protein with available structure information exists among the ThDP-dependent enzymes. As the overall structure of a homotrimeric PPDC is expected to deviate considerably from the dimeric structures of the known ThDP-dependent enzymes, a correlation between linker length and intramonomer architecture cannot be assumed.

SPDC: The observation of dimeric and trimeric forms of PPDCs implies the possibility of two different architectures for the homologous $\alpha_6\beta_6$ -heterododecameric SPDCs, either resulting from triplication of an $\alpha_2\beta_2$ -heterotetramer or from duplication of an $\alpha_3\beta_3$ -heterohexamer. The $\alpha_2\beta_2$ -heterotetrameric and the $\alpha_3\beta_3$ -heterohexameric SPDCs are expected to correspond to the homodimeric and the homotrimeric PPDCs, respectively.

Evolutionary relationships of ThDP-dependent enzymes

Previous investigations on the sequence relationships and the structural similarity of ThDP-dependent enzymes from different superfamilies led to possible evolutionary pathways from a common ancestor to contemporary ThDP-dependent enzymes (Costelloe, Ward, and Dalby 2008; Duggleby 2006). In this work, we analyzed the similarity of a large number of PYR and PP domains confirming the previous findings in general.

As proposed by Todd et al. (Todd, Orenge, and Thornton 2001), the PYR and PP domains have been evolved from a common ancestor and subsequently been optimized for the binding of the ThDP cofactor from the PYR and the PP site, respectively (Figure 4.8 on page 115). Based on structure and sequence analysis, this putative protein served as the common ancestor ('PYR/PP-ancestor') for two evolutionary branches: one branch leading to SPDCs, PPDCs, and DCs, and a second branch leading, after recruitment of a TKC domain, to TKs, DXPSs, aKGDHs, aKADH1s, and aKADH2s. In the previous investigations of the evolution of ThDP-dependent enzymes, SPDCs and PPDCs were assumed to subsequently be evolved from the PYR/PP-ancestor (Costelloe, Ward, and Dalby 2008; Duggleby 2006). The SPDCs were supposed to

have evolved by duplication and subsequent uncoupled evolution into an $\alpha_2\beta_2$ -heterotetrameric protein and, in a next step, both domains were assumed to be coupled into a single protein chain to yield PPDCs. However, gel filtration chromatography revealed an $\alpha_6\beta_6$ -dodecameric protein (Graupner, Xu, and White 2000; Denger et al. 2009). To explain the evolution of an homotrimeric PPDC as found in *B. fragilis* (Zhang, Dai, et al. 2003), an $\alpha_3\beta_3$ -heterohexamer has to be assumed as an ancestor.

Thus, for the branch leading to SPDCs, PPDCs, and DCs, we propose an SPDC/PPDC-ancestor capable to form both $\alpha_2\beta_2$ -heterotetrameric and $\alpha_3\beta_3$ -heterohexameric complexes. Based on the high sequence similarity of SPDCs and PPDCs to DCs, their similar structural architecture, and the lower sequence similarity to other superfamilies, the DCs probably have evolved from the $\alpha_2\beta_2$ -heterotetrameric SPDC/PPDC-ancestor by recruitment of the TH3 domain and the linkage of the PYR and PP domains. In parallel, by linkage of the PYR and PP domains via a short linker without inclusion of any additional domain, the homodimeric PPDCs, as found in *S. viridochromogenes* (Johnen and Sprenger 2009), have been evolved. The observed $\alpha_6\beta_6$ -dodecameric structure of SPDCs could be explained by triplication of the $\alpha_2\beta_2$ -heterotetrameric SPDC/PPDC-ancestor or duplication of an ancestral $\alpha_3\beta_3$ -heterohexameric SPDC/PPDC protein, which after linkage of the PYR and PP domains resulted in a homotrimeric PPDC.

The development of the TKC-containing superfamilies is most probably linked to a common heterotetrameric TKC ancestor after recruitment of the TKC domain (Costelloe, Ward, and Dalby 2008). The OR superfamily shows only weak similarity to the other TKC-containing superfamilies. In the networks of the PYR and PP domains, the ORs are more similar to the DCs than to TKs or aKADH2s. In an evolutionary context, this would result from early separation of the OR superfamily from the other TKC-containing enzymes and homologous development of ORs and DCs (Costelloe, Ward, and Dalby 2008). Although structural information is available only for the PFOR family, sequence analysis allows insights into the evolutionary diversification within the OR superfamily. While the PYR and the PP domains are located on different chains in VOR and KGOR, both domains are connected via additional domains in the majority of PFOR and IOR. After recruitment of the domains III/IV and V, the PYR-TKC complex was linked to the PP domain in PFOR and IOR. In addition, domain VII was C-terminally linked to the PP domain in both families. Thus, the VORs, KGORs, and the minority of PFORs and IORs having the PYR and PP domains on separate protein chains have probably evolved from intermediate proteins on the evolutionary path to the complex proteins having six domains fused into one protein chain.

Among the subfamily of TKC-containing ThDP-dependent enzymes, the domain structure of the intra-monomer/PYR-PP type of fused ORs is unique. All other TKC-containing enzymes, except for the $\alpha_2\beta_2$ -heterodimeric aKADH2s, have their catalytic domains fused in the opposite PP-PYR order. Thus, in parallel to the evolutionary separation toward the PYR-PP type OR superfamily, the homodimeric TKC ancestor developed toward the PP-PYR type superfamilies TK, aKGDH, and aKADH1.

Subsequently, the intra-monomer/PP-PYR type DXPSs were evolved from the TKs by shortening the linker between the PP and the PYR domain (Costelloe, Ward, and Dalby 2008), which led to domain rearrangement and thus resulted in a switch from an inter-monomer to an intra-monomer type. The high global sequence similarity of TKs and DXPSs demonstrates that this event must have happened recently. Besides a reduction of length, a complete deletion of the linker occurred in 38% of all TK sequences, which have their PP domain and a fusion of the PYR and the TKC domain on separate chains. In contrast, the suggested evolutionary development of the inter-monomer/PP-PYR type TKs by circular permutation of a PYR-PP type (Schenk, Duggleby, and Nixon 1998) is less probable, since both PYR-PP type superfamilies, the DCs and the ORs, have additional domains that are lacking in the TK superfamily. The 2-oxoglutarate dehydrogenase family (OGDH) of the aKGDH superfamily evolved from the homodimeric TKC ancestor maintaining the structural architecture, and the 2-oxoglutarate decarboxylase family (OGDC) evolved by recruitment of an additional C-terminal ATN domain.

Previously, the ATN domain was described to be a sequence feature, which is only found in proteins from corynebacterineae (Wagner et al. 2011). By analyzing the N-terminal sequences of 2219 homologous OGDC sequences, this domain was also found in further suborders of actinobacteria like streptomycineae, streptosporangineae, micrococcineae, pseudonocardineae, propionibacterineae, glycomycineae, frankineae, and others.

Notably, the aKADH2 superfamily is the only directly connected family to the aKGDH superfamily in the similarity networks of the PYR and PP domains. In former investigations, the E1 component of the 2-oxoglutarate dehydrogenase (OGDH) was grouped into the aKADH2 superfamily (Duggleby 2006), which was confirmed by the connection of both families in the networks of the PYR and PP domain similarities. However, based on the different structural architectures of aKADH2s and aKGDHs and the high global sequence similarity between the two homologous families of aKGDHs (OGDH and OGDC), both families, the aKADH2s and aKGDHs, were

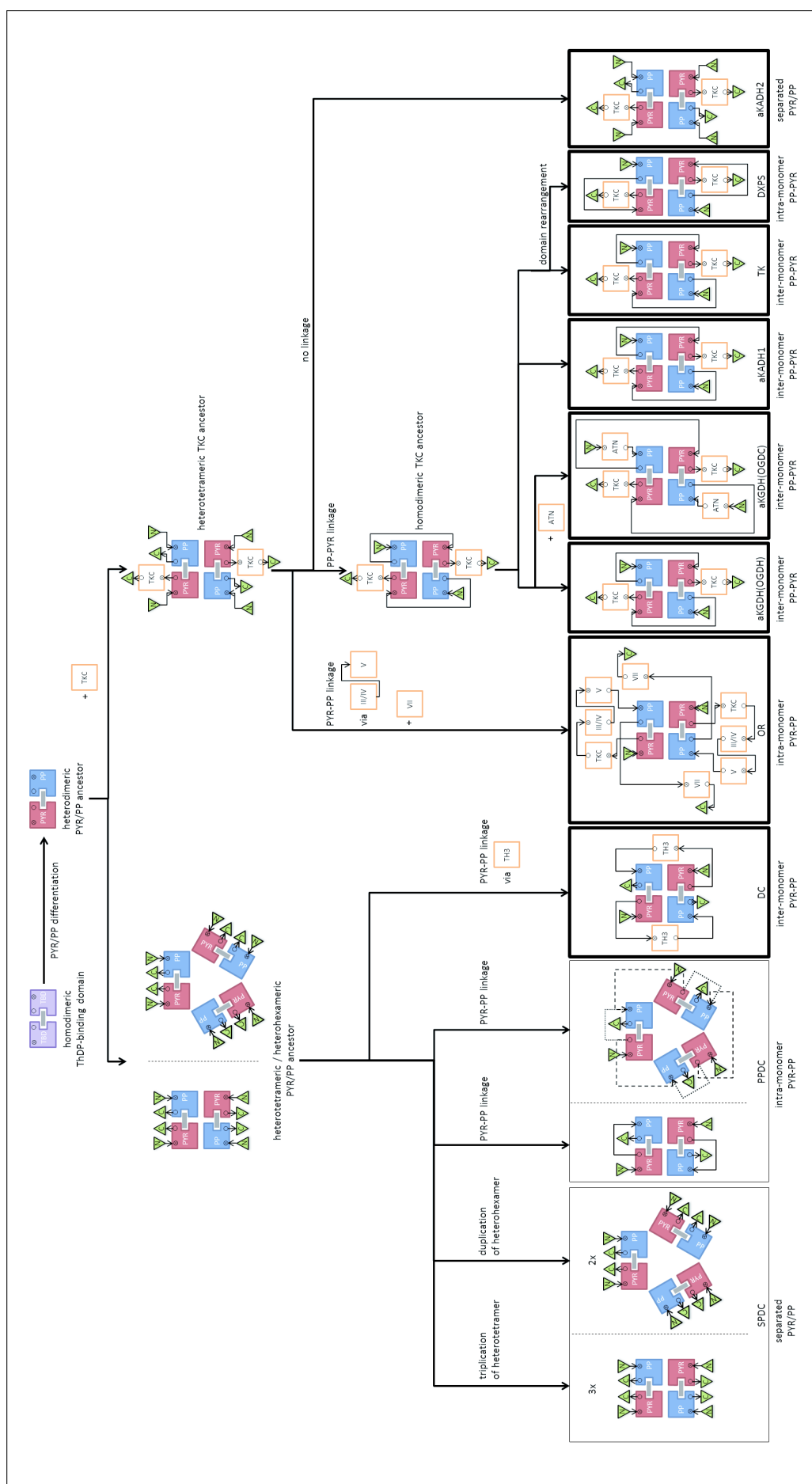


Figure 4.8: Proposed evolutionary pathway of ThDP-dependent enzymes. The PYR and PP domains are supposed to have evolved from a common ancestor (Todd, Orengo, and Thornton 2001) and have subsequently developed into nine superfamilies. The structurally different OGDC and OGDH families were based on sequence similarity both assigned to the aKGDH superfamily. Thick bordered boxes represent superfamilies with structure information. For SPDC and PPDC, the structure is not known. The two theoretically possible architectures of homotrimeric PPDC (intra-monomer and inter-monomer) are represented by dotted and dashed lines, respectively. By recruitment of additional domains, linkage of the catalytic domains in different order and domain rearrangement, ThDP-dependent enzymes have evolved to a structurally diverse protein family.

separated into two distinct superfamilies. The aKADH1 and aKADH2 superfamilies have been previously regarded to be evolutionary closely linked (Duggleby 2006). However, due to their moderate sequence similarity, their deviating structural composition of the PP domain, and the difference in wiring of PP and PYR domains, we assume that they have been separated early in evolution into two independent evolutionary branches. From a structural point of view, the aKADH2 superfamily could have been evolved from the heterotetrameric TKC ancestor before the linkage of the catalytic domains led to the homodimeric TKC ancestor of TKs, DXPSs, aKGDHs, aKADH1s, and ORs. However, based on sequence similarity of the PP and PYR domains, the aKADH2 superfamily is highly connected to other TKC-containing enzymes. By sequence similarity of the PP domains, the aKADH2 superfamily is even highly connected to the DC superfamily.

Thus, the PP domains of ThDP-dependent enzymes from different superfamilies show a remarkable sequence similarity despite deviating in function and overall structure. Although sequence similarity is lower between the PYR domains, the basic structure of both networks is similar indicating coupled evolution of the catalytic domains. This is supported by analyzing their relative sequence similarity as derived from PYR-PP and PP-PYR type fusion proteins, where average sequence similarities of PP and PYR domains were highly correlated (Figure 4.7 on page 109). Based on the proposed evolutionary route to contemporary ThDP-dependent enzymes, the role of the additionally recruited domains such as TH3, TKC, ATN, or the OR specific domains is still an open question. The most obvious architectural difference between the SPDCs and PPDCs and all other superfamilies is the lack of additional domains. Therefore, one effect of the additionally acquired domains might be to hinder the formation of multimers other than dimers and to stabilize the quaternary structure. In addition, the TKC domain in aKADH2 enzymes seems to participate in the recognition of the peripheral subunit binding domain of the E2 component in the multienzyme complex (Frank, Pratap, et al. 2005). However, removal of the TKC-domain attached to the PYR domain of the transketolase from *E. coli* resulted in an even increased catalytic activity, eliminating any functional effect for this enzyme (Costelloe, Ward, and Dalby 2008). Furthermore, the TH3 domain of enzymes from the DC superfamily has lost its original effect of binding FAD in most homologous families of DCs. From the entire superfamily, only pyruvate oxidase is known to require FAD for proper function (Tittmann, Wille, et al. 2005), and for the oxalyl-CoA decarboxylase (OCDC) an evolutionary adjustment of the former FAD binding site to an ADP binding site was demonstrated (Berthold, Moussatche, et al. 2005). In the

enzymes glyoxylate carboligases (GXC) (Chung, Tan, and Suzuku 1971), cyclohexane-1,2-dione hydrolase (CDH) (Steinbach et al. 2012), and acetohydroxy acid synthases (AHAS) (Lee, Lee, et al. 2013; Duggleby and Pang 2000; McCourt et al. 2006), FAD is still bound in the active enzyme but no functional role has been observed, yet. This supports the assumption of a predominantly structural role of additional domains. As an exception, the ATN domain of OGDC from the aKGDH superfamily was shown to act in allosteric activation of the respective enzymes (Wagner et al. 2011).

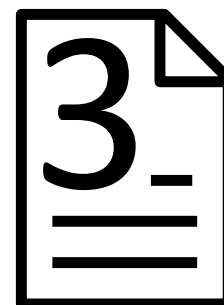
4.2.6 Conclusions

The ThDP-dependent enzymes family encompasses more than 50000 different proteins with nearly 80000 sequences. Large-scale analysis allowed for the examination of the family structure and the relationships between 168 homologous families and 9 superfamilies. The homologous families within the superfamilies all show moderate to high global sequence similarity, but due to the different sequential order of the PYR and the PP domain, global sequence similarity is not an adequate measure for sequence relationships between the superfamilies. Based on representative structures for seven superfamilies, a comprehensive overview over the structural diversity became feasible. Moreover, a profile-based domain identification allowed for the detection of the structurally conserved PYR and PP domains in the majority of ThDP-dependent enzymes and thus for the analysis of the sequential relationship on a domain level. The evolutionary separation into different superfamilies was explained by a classification into five different types of structural architectures. For the TKC-containing families, the sequential PP-PYR order is the most common, whereas members of the DC and PPDC superfamilies share the PYR-PP order. Furthermore, the majority of the TKC-containing enzymes is represented by the inter-monomer type, except for ORs, which resulted from a separate evolutionary path, and DXPSs which resulted from a domain rearrangement of TKs. While recruitment of the TH3 domain yielded an inter-monomer type for DCs, the PPDCs are assumed to have evolved into an intra-monomer type. SPDCs and aKADH2s, having the PYR and PP domains on separated protein chains, are representing the structures of the early ancestors. ThDP-dependent enzymes are highly diverse in their global sequence and in their global structure, but have highly conserved active sites composed by the PYR and PP domains. Deepening our insight into the modular architecture of proteins is crucial for a more successful engineering and design of novel biocatalysts. PYR

and PP domains from different ThDP-dependent enzymes can be seen as modules that can be recombined to change substrate specificity, chemo- and stereoselectivity, and thus to enlarge the toolbox of industrially desired enzymes for selective C-C bond ligation or cleavage including reactions that are exclusively accessible by N-heterocyclic carbene organocatalysis, but not yet by ThDP-dependent enzymes.

4.3 A tailor-made chimeric thiamine diphosphate-dependent enzyme for the direct asymmetric synthesis of (*S*)-benzoins

Westphal, R; Vogel, C; Schmitz, C; Pleiss, J; Müller, M; Pohl, M; Rother, D. (2014) A tailor-made chimeric thiamine diphosphate-dependent enzyme for the direct asymmetric synthesis of (*S*)-benzoins. *Angewandte Chemie International Edition* 53 (35): 9376-9379



4.3.1 Abstract

Thiamine diphosphate-dependent enzymes are well known for catalyzing the asymmetric synthesis of chiral α -hydroxy ketones from simple prochiral substrates. The steric and chemical properties of the enzyme active site define the product spectrum. Enzymes catalyzing the carbonylation of aromatic aldehydes to (*S*)-benzoins have not so far been identified. We were able to close this gap by constructing a chimeric enzyme, which catalyzes the synthesis of various (*S*)-benzoins with excellent enantiomeric excess (>99%) and very good conversion.

4.3.2 Communication

Thiamine diphosphate (ThDP) dependent enzymes are proven catalysts for the stereo- and regioselective synthesis of chiral α -hydroxy ketones through asymmetric C-C bond formation starting from aldehydes, α -keto acids, and ketones (Hoyos et al. 2010; Brovetto et al. 2011). Whereas enzymes are available for the enantio-complementary synthesis of a broad range of symmetrical and mixed aliphatic, araliphatic, and aromatic products (Müller, Gocke, and Pohl 2009; Müller, Sprenger, and Pohl 2013; Hailes et al. 2013), so far no biocatalyst is available to selectively access (*S*)-benzoins starting from benzaldehydes. Instead, other enzymatic routes became available, which involve kinetic resolution as well as deracemization of *rac*-benzoins (Demir, Pohl, et al. 2001; Fragnello et al. 2012), or asymmetric reduction of 1,2-diarylethane-1,2-diones (benzils) (Demir, Peruze, et al. 2008). Although these approaches are mostly characterized by high stereoselectivity, they all share the same drawback: the racemic or prochiral starting

material must be chemically synthesized beforehand. Such additional steps reduce the eco-efficiency and sustainability of the process compared to the direct enzymatic synthesis of (*S*)-benzoin starting from commercially available benzaldehydes. This direct synthesis is still a unique feature of nonenzymatic synthetic organic chemistry (Enders and Kallfass 2002b; Enders and Kallfass 2002a). Whereas the *R*-selective enzymatic synthesis of benzoin has already been achieved (Demir, Dünnwald, et al. 1999; Demir, Sesenoglu, Eren, et al. 2002; Dünkemann et al. 2002), (*S*)-selective synthesis is limited by the specific active-site architecture of ThDP-dependent enzymes (Figure 4.9).

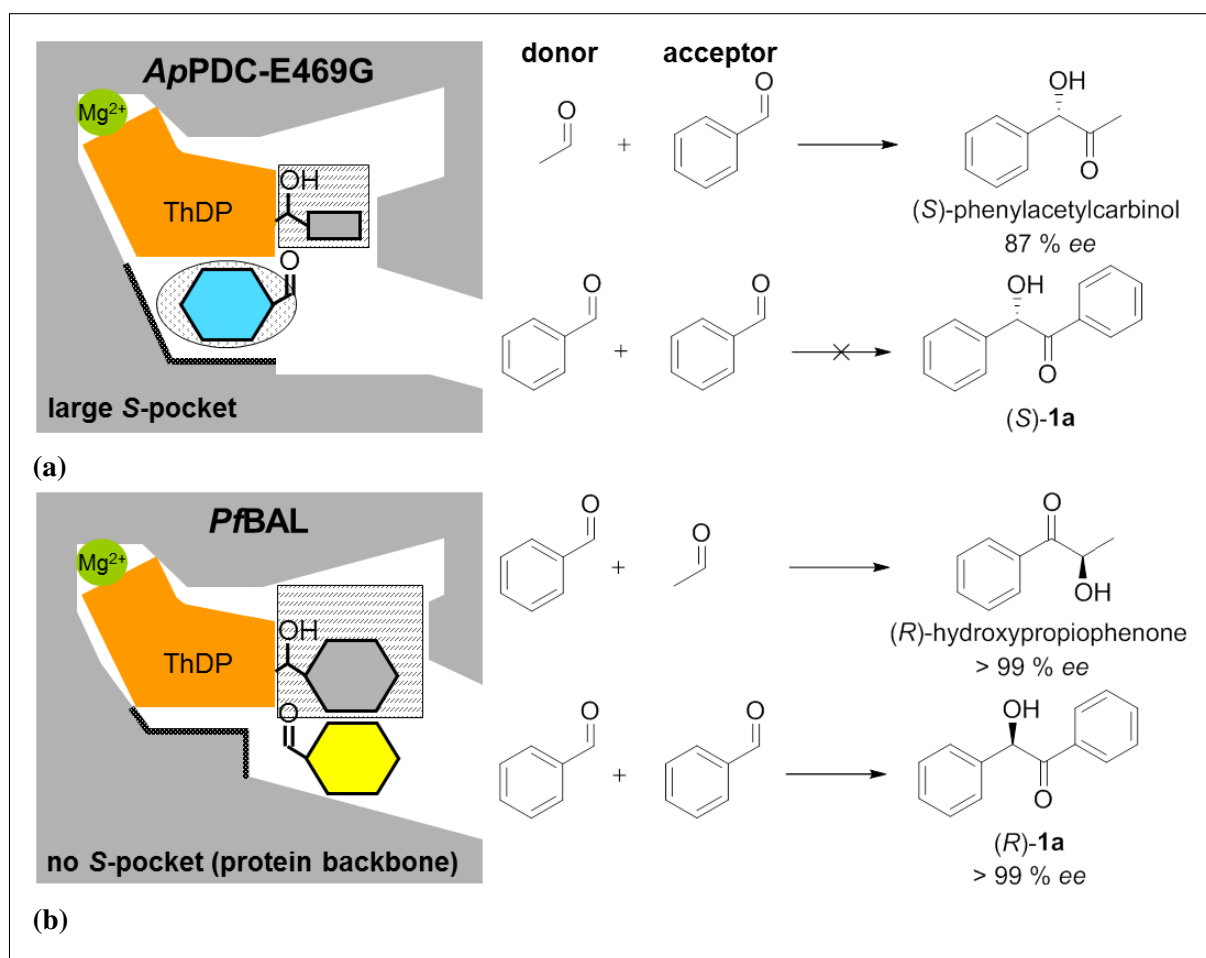


Figure 4.9: A schematic presentation of the scope and limitations of carboligations catalyzed by *ApPDC-E469G* and *PfbAL*. (*S*)-Benzoin (**1a**) formation is prevented either by a small donor-binding site (a), or an *S*-pocket that is inaccessible for the antiparallel orientation of the acceptor benzaldehyde relative to the ThDP-bound donor benzaldehyde (as a hydroxybenzyl group; b). Dashed rectangle: donor-binding site; dotted circle: *S*-pocket region; rectangle: acetaldehyde; hexagon: benzaldehyde; gray: donor; cyan: antiparallel-oriented acceptor, which leads to the respective (*S*)-product; yellow: parallel-oriented acceptor, which leads to the respective (*R*)-product.

Introduction of the *S*-pocket concept enabled the design of an (*S*)-selective variant of pyruvate decarboxylase from *Acetobacter pasteurianus* (*ApPDC*) for the formation of (*S*)-phenylacetylcarbinol (PAC) (Rother et al. 2011). Replacement of a glutamate residue (E469) by glycine allowed

benzaldehyde as an acceptor to bind predominantly in antiparallel orientation relative to the ThDP-bound donor, which is a prerequisite for (*S*)-selectivity. However, *ApPDC*-E469G only provides a small donor-binding site that preferentially stabilizes small aliphatic donor substrates, thereby preventing the formation of (*S*)-benzoin [(*S*)-**1a**] (Figure 4.9 a). Benzaldehyde lyase from *Pseudomonas fluorescens* (*PfBAL*) is a powerful but strictly *R*-selective catalyst for the synthesis of **1a** (Demir, Sesenoglu, Eren, et al. 2002). In contrast to *ApPDC*, *PfBAL* provides a large donor-binding site, which is ideal for stabilizing benzaldehyde in that position. However, owing to a missing *S*-pocket, *PfBAL* only allows the parallel arrangement of the donor and acceptor benzaldehydes prior to carbonylation. Furthermore, *S*-pocket engineering is limited in *PfBAL* by the position of the protein backbone of the respective α -helix in the *S*-pocket region (Figure 4.9 b). To solve the long-standing problem of enzymatic (*S*)-benzoin synthesis, a rational hybridization approach was followed, in which the active-site characteristics of the variant *ApPDC*-E469G (Rother et al. 2011) and *PfBAL* were combined. Two approaches were conceivable:

1. the introduction of a large *S*-pocket into *PfBAL*, or
2. the extension of the donor-binding site of *ApPDC*-E469G

Herein, we show that the second approach indeed resulted in the first tailor-made (*S*)-selective ThDP-dependent enzyme variant for the formation of various benzoins. A combination of modeling studies and a comprehensive sequence analysis of the amino acid distribution in 186 homologous PDC sequences and 43 BAL sequences led to the identification of a threonine residue (T384) as the pivotal factor that influences the size of the donor-binding site in *ApPDC*-E469G. T384 limits the space for benzaldehyde in the donor-binding site and is conserved in 92% of *ApPDC* homologous sequences (Figure 4.10 b on page 122), whereas glycine was found at the equivalent position in all *PfBAL* homologous sequences. T384 in *ApPDC*-E469G was replaced with glycine through site-directed mutagenesis in order to mimic the *PfBAL* donor-binding site. As a result, an enlarged donor-binding site with putatively sufficient space for benzaldehyde was obtained (Figure 4.10 b). Moreover, short molecular dynamics simulations revealed that a neighboring tryptophan residue (W388) underwent minor conformational changes, probably because of interactions with the ThDP-bound benzaldehyde donor (hydroxybenzyl), thereby further opening up the donor-binding site.

Biochemical characterization of the new variant proved that the single mutation of T384 to

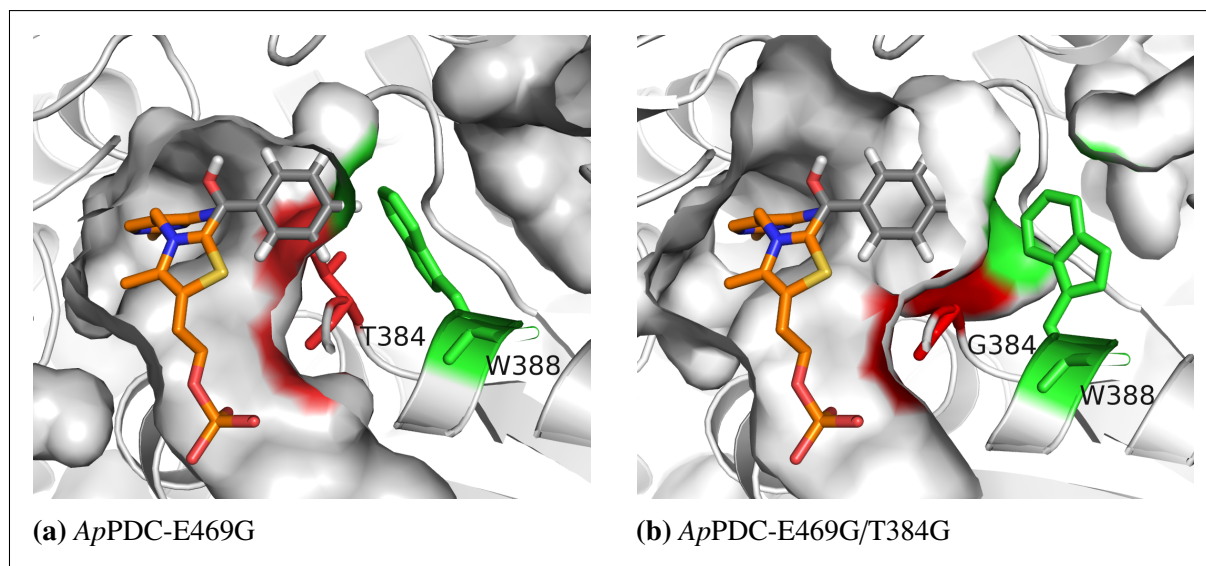


Figure 4.10: The donor-binding sites of *ApPDC*-E469G (a) and *ApPDC*-E469G/T384G (b) with benzaldehyde bound as a hydroxybenzyl group (gray) to C2 of ThDP (orange). a) *ApPDC*-E469G is not able to properly bind benzaldehyde in the small donor-binding site, which is mainly restricted by T384 (red). b) The donor-binding site could be opened for benzaldehyde by replacing T384 with glycine. Moreover, the equilibrated structure revealed conformational changes to W388 (green) that additionally opened the donor-binding site.

glycine indeed altered the chemoselectivity of *ApPDC*-E469G, which subsequently preferred benzaldehyde as the donor. Whereas no significant formation of **1a** in the homocoupling of benzaldehyde was observed with *ApPDC*-E469G (Table 4.4, entry 1), double variant *ApPDC*-E469G/T384G catalyzed the benzoin formation with 52% conversion under the tested conditions (entry 2).

Table 4.4: Enzymatic synthesis of (*S*)-**1a** as catalyzed by *ApPDC* variants^[a].

Entry	<i>ApPDC</i> variant	ee [%] ^[b]	Conv. [%] ^[c]
1	E469G	n.d. ^[d]	<1
2	E469G/T384G	59	52
3	E469G/T384G/I468G	66	23
4	E469G/T384G/I468A	87	95
5	E469G/T384G/I468V	76	40
6	E469G/T384G/I468A/W543F	95	36

^[a] Reaction conditions: 50 mM triethanolamine buffer, pH 8.0, 2 mM MgSO₄, and 0.1 mM ThDP; 1 mg mL⁻¹ enzyme; 18 mM benzaldehyde; 20 °C, 6 h

^[b] Determined by chiral-phase HPLC.

^[c] Determined by chiral-phase HPLC based on the consumption of benzaldehyde.

^[d] Not determined.

The preference for benzaldehyde as the donor was also demonstrated in the mixed carbonylation of acetaldehyde and benzaldehyde by variant *ApPDC*-E469G/T384G, which resulted in the

formation of a mixture of **1a** and 2-hydroxypropiophenone (2-HPP), whereas PAC, the main product obtained with wild-type *Ap*PDC or variant *Ap*PDC-E469G (Rother et al. 2011), was only detected in trace amounts. Furthermore, variant *Ap*PDC-E469G/T384G catalyzed both the synthesis of (*S*)-**1a** from benzaldehyde with moderate stereoselectivity (59% ee, entry 2) and mixed carboligation towards 2-HPP with good (*S*)-selectivity (91% ee). To improve the moderate (*S*)-selectivity of *Ap*PDC-E469G/T384G, two strategies are possible:

- stabilization of the antiparallel ('*S*-pathway') or
- destabilization of the parallel acceptor orientation ('*R*-pathway') prior to carboligation (Westphal, Hahn, et al. 2013).

To suppress the *R*-pathway in *Ap*PDC-E469G/T384G, the active site was examined for residues that potentially stabilize the acceptor benzaldehyde in the parallel orientation. By performing molecular modeling, I468 and W543 were identified (Figure 4.11 a) as residues that could stabilize parallel-oriented benzaldehyde through nonpolar interactions or π -stacking. I468 was replaced with valine, glycine, or alanine to increase the distance of the respective side chain to parallel-oriented benzaldehyde and thus disrupt the stabilization of the *R*-pathway (Figure 4.11 b).

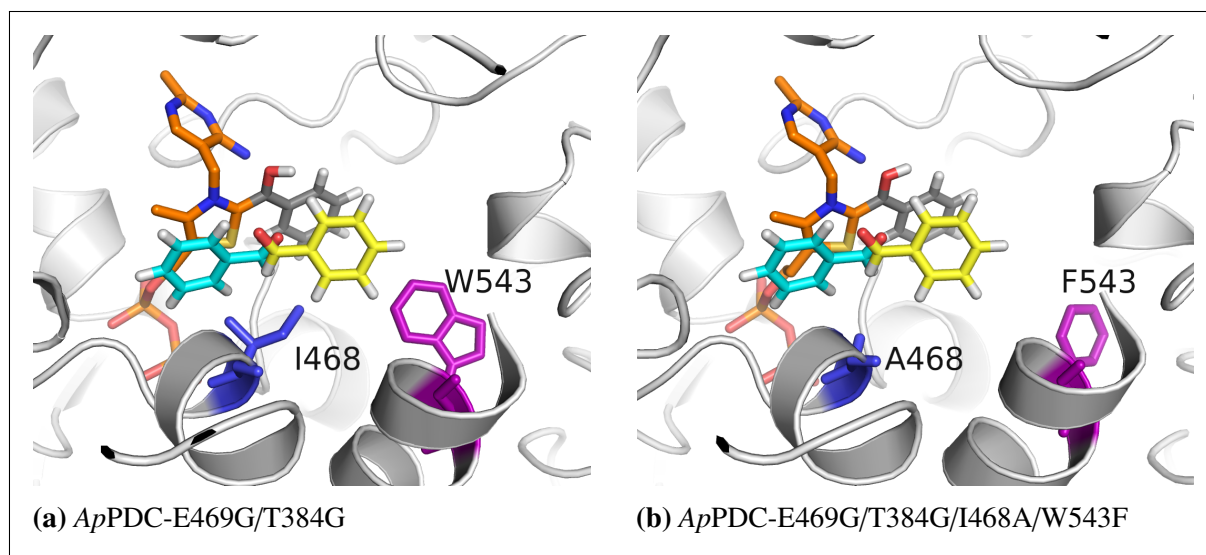


Figure 4.11: Possible stabilization of the parallel-oriented acceptor benzaldehyde (yellow) in the active sites of *Ap*PDC-E469G/T384G (a) and *Ap*PDC-E469G/T384G/I468A/W543F (b). a) In E469G/T384G, parallel-oriented benzaldehyde is putatively stabilized by I468 (blue) and W543 (purple), neither of which directly influence its antiparallel orientation (cyan). b) In variant E469G/T384G/I468A/W543F, the stabilization of parallel-oriented benzaldehyde by A468 (blue) and F543 (purple) is no longer possible.

In all cases, the new variants showed improved (*S*)-selectivity for the formation of (*S*)-**1a** (Table 4.4, entries 3-5). Out of the different variants, the triple variant E469G/T384G/I468A

performed best with respect to *ee* (87%, *S*) and conversion (95%) under standard reaction conditions. The high conversion is particularly surprising because multiple active-site mutations frequently result in a drastic decrease in enzymatic activity, as has been demonstrated for other ThDP-dependent enzymes (Westphal, Hahn, et al. 2013). To further improve the stereoselectivity of variant E469G/T384G/I468A, a fourth mutation at the position of tryptophan 543 was introduced. W543 is part of the C-terminal α -helix, which covers the entrance to the active site of *ApPDC* (Figure 4.11 a), and of an aromatic cluster (for details, see Supporting Information, Section A.6.2 on page 217) that might be relevant for structural stabilization. This position was subjected to site-saturation mutagenesis by using NDT codon degeneracy. Only few variants showed significant carboligation activity during screening, but all of the active variants were (*S*)-selective for the formation of **1a**. Among these variants, E469G/T384G/I468A/W543F revealed improved (*S*)-selectivity for the synthesis of **1a**, with 95% *ee* and a conversion of 36% (Table 4.4, entry 6). In comparison to W543, the increased distance between the phenyl ring of the parallel-oriented acceptor benzaldehyde and F543 (6.8 Å compared to 3.7 Å) is assumed to prevent stabilizing interactions (Figure 4.11 b). Furthermore, the physicochemical properties at position 543 were preserved by the substitution of tryptophan by phenylalanine, which might be beneficial to maintaining the structural stability conferred by the aromatic cluster in the *ApPDC* variant. In addition to improved (*S*)-selectivity in the homocoupling reaction of benzaldehyde, *ApPDC*-E469G/T384G/I468A/W543F also revealed enhanced stereoselectivity for the synthesis of (*S*)-2-HPP (>99% *ee*) starting from benzaldehyde as the donor and acetaldehyde as the acceptor substrate. To evaluate the synthetic potential of the newly designed variants, a substrate screening with substituted benzaldehyde derivatives was performed with *ApPDC*-E469G/T384G/I468A and *ApPDC*-E469G/T384G/I468A/W543F (Table 4.5 on page 125) under optimized reaction conditions (reduced temperature of 15 °C) for the synthesis of (*S*)-**1a**. *meta*-Substituted benzaldehyde derivatives turned out to be the best substrates in the homocoupling reaction in terms of (*S*)-selectivity and conversion. Except for 3-fluorobenzaldehyde (**1e**), all of the *meta*-substituted benzaldehydes (Table 4.5, entries 5-9) were transformed with higher (*S*)-selectivity than benzaldehyde (Table 4.5, entry 1). The reactions with variant E469G/T384G/I468A/W543F again showed higher (*S*)-selectivity and lower conversion than those with variant E469G/T384G/I468A. Remarkably, variant E469G/T384G/I468A/W543F catalyzed the synthesis of enantiopure **1f-i** (>99% *ee* (*S*), Table 4.5, entries 6-9). These synthesis reactions could be successfully scaled-up from analytical scale (300 μ L) to preparative scale (20 mL) with final product concentrations of up

to 4 gL⁻¹ (yields of isolated product 61-85%, Table 4.5). In contrast to *meta*-substituted benzaldehydes, *ortho*- and *para*-substituted benzaldehydes, with the exception of 4-fluorobenzaldehyde (**1j**), were converted with low conversion to yield an excess of the *R*-enantiomer.

Table 4.5: Synthesis of benzoins (**1a-l**) as catalyzed by ApPDC variants^[a].

Entry	Ar	Product	E469G/T384G/I468A		E469G/T384G/ I468A/W543F	
			ee [%] ^[b]	Conv. [%] ^[c] (yield [%]) ^[d]	ee [%] ^[b]	Conv. [%] ^[c] (yield [%]) ^[d]
1	C ₆ H ₅	1a	89 (<i>S</i>)	92 (85)	98 (<i>S</i>)	26 (66)
2	2-FC ₆ H ₄	1b	58 (<i>R</i>)	36	21 (<i>R</i>)	<5
3	2-ClC ₆ H ₄	1c	75 (<i>R</i>)	<5	n.d.	n.c.
4	2-MeOC ₆ H ₄	1d	n.d.	n.c.	n.d.	n.c.
5	3-FC ₆ H ₄	1e	87 (<i>S</i>)	80 (82)	93 (<i>S</i>)	36
6	3-ClC ₆ H ₄	1f	91 (<i>S</i>)	97	>99 (<i>S</i>)	48 (72)
7	3-BrOC ₆ H ₄	1g	95 (<i>S</i>)	85 (84)	>99 (<i>S</i>)	30
8	3-IC ₆ H ₄	1h	96 (<i>S</i>)	30 (70)	>99 (<i>S</i>)	11
9	3-MeOC ₆ H ₄	1i	98 (<i>S</i>)	93	>99 (<i>S</i>)	58 (61)
10	4-FC ₆ H ₄	1j	64 (<i>S</i>)	53	85 (<i>S</i>)	<5
11	4-ClC ₆ H ₄	1k	>99 (<i>R</i>)	10	n.d.	n.c.
12	4-MeOC ₆ H ₄	1l	n.d.	<5	n.d.	n.c.

^[a] Reaction conditions: see Table 4.4 on page 122, reaction temperature: 15 °C

^{[b][c]} See Table 4.4

^[d] Yields of isolated product after preparative synthesis and conversion of >90% after 24 h and 48 h, (reaction conditions: see Supporting Information, Section A.6.4).

^[e] Not determined.

^[f] No conversion.

These results confirm our findings from recent studies on 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase (MenD), which revealed that *meta*-substituted benzaldehydes are converted with higher (*S*)-selectivity and conversion than *ortho*- and *para*-substituted benzaldehydes and unsubstituted benzaldehyde (Westphal, Hahn, et al. 2013; Westphal, Jansen, et al. 2014). However, the reasons for the preference for *meta*-substituted benzaldehydes in (*S*)-selective carbonylation reactions and the switch in stereoselectivity in the case of *ortho*- and *para*-substituted benzaldehydes are still not understood. In this case, a combination of steric and electronic interactions between the substrate and the *S*-pocket might influence the stereoselectivity. The elucidation of this phenomenon is now the subject of further investigation. In summary, the rational design of a hybrid substrate-binding site, which combines the active-site characteristics of the (*S*)-selective ApPDC-E469G variant and PfBAL, enabled the creation of a new biocatalyst for the synthesis of different (*S*)-benzoins with excellent ee values and good

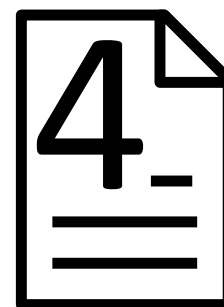
conversion when starting from commercially available benzaldehydes. These results highlight the robustness of ThDP-dependent enzymes with respect to active-site mutations, as well as their tremendous catalytic potential in asymmetric carbonylation reactions. Moreover, this hybridization approach might open the field for the combinatorial assembly of selectivity-determining modules from different ThDP-dependent enzymes and offer a new perspective on thiamine catalysis with respect to the design of variants with novel catalytic activities.

4.3.3 Experimental Section

All chemicals were purchased from Sigma Aldrich. Benzaldehydes were freshly distilled before use. The generation, expression, and purification of ApPDC variants are described in the Supporting Information (see Section A.6 on pages 216ff.). Descriptions of sequence and structural analyses, as well as reaction details and product analytics, can also be found in the Supporting Information.

4.4 BioCatNet: a system for the analysis of sequence-structure-function relationships of protein families

Vogel, C; Reusch, W; Rother, D; Pohl, M; Pleiss, J. BioCatNet: a system for the analysis of sequence-structure-function relationships of protein families. *Manuscript in preparation.*



4.4.1 Abstract

BioCatNet is an infrastructure for the generation, maintenance and analysis of family-specific protein databases, which have extensively been applied during the last decade to explore family structures and to support rational protein engineering. Rational protein engineering is an efficient and minimally invasive strategy to improve catalytic activity, stability, chemo-, regio- or stereoselectivity. However, it requires detailed information on the general and specific sequence-structure-function relationships of the respective enzyme family under investigation. Family-specific protein databases provide comprehensive information on protein sequences and structures, functional annotations, the effects of mutations, as well as links to literature, and thus provide a basis for systematic analyses. However, publicly available family-specific protein databases are focused to sequence and structure information and neglect detailed information on the biochemical properties of the respective biocatalysts. Moreover, if biochemical data is provided in the literature, an unambiguous link to sequence information is often lacking, which hinders the systematic analysis. *BioCatNet* was developed to close this gap; sequence and structure information is gathered from publicly available repositories by automated processes, a manual strategy was chosen for the acquisition of biochemical data by providing an electronic laboratory journal. As a consequence, *BioCatNet* provides access to the large resources of biochemical data that accumulate in experimentally working laboratories by integrating heterogeneous data from biocatalytic experiments into a unified data model. Herein we report on the concept and the implementation of the *BioCatNet* system and demonstrate its applicability.

As a first example, the ‘Thiamine diphosphate-dependent Enzyme Engineering Database’ was migrated to *BioCatNet* (www.teed.biocatnet.de).

4.4.2 Background

Enzyme engineering based on data mining and molecular modeling has become a promising time- and cost-efficient strategy for the development of desired enzymes variants. A prerequisite for rational engineering is detailed knowledge about the relationship between sequence and structure of an enzyme and its biochemical properties (Arnold 2001). Consequently, information on protein sequences and structures and on the effects of mutations on biochemical properties have become precious raw materials for successful protein design approaches. Comprehensive online repositories such as the NCBI Protein Database (Benson et al. 2011), UniProt (The UniProt Consortium 2014), DDBJ (Kaminuma et al. 2011), and Protein Data Bank (Berman et al. 2000) enable systematic analyses of sequences and structures of various biocatalysts. To unravel the functionally most relevant amino acids or to identify promising mutation sites, positions were identified that are conserved (Lehmann et al. 2002; Anbar et al. 2012), specific for a family of homologous proteins (Mazin et al. 2010; Suplatov, Besenmatter, et al. 2012; Suplatov, Panin, et al. 2014), or correlated within a homologous family (Kuipers et al. 2009; Kowarsch et al. 2010). Therefore, family-specific protein databases have been established for a large number of enzymes such as cytochrome P450 monooxygenases (Nelson 2009; Fischer, Knoll, et al. 2007; Gricman, Vogel, and Pleiss 2014; Sirim, Wagner, Lisitsa, et al. 2009; Sirim, Widmann, et al. 2010), β -lactamases (Widmann, Pleiss, and Oelschlaeger 2012), triterpene-cyclases (Racolta et al. 2012), imine reductases (Scheller et al. 2014), lipases and other α/β -hydrolases (Barth et al. 2004a; Barth et al. 2004b; Pleiss et al. 2000; Widmann, Juhl, and Pleiss 2010; Fischer and Pleiss 2003; Fischer, Thai, et al. 2006; Hotelier et al. 2004), carbohydrate-binding enzymes (Lombard et al. 2014), and thiamine diphosphate-dependent enzymes (Widmann, Radloff, and Pleiss 2010). Although existing family-specific protein databases make a rapidly increasing number of sequences and structures of proteins accessible to systematic analysis, they lack complete information on the biochemical properties of the respective enzymes. Providing stringently linked sequences, structures and details on biochemical properties of enzymes would result in an invaluable information resource for deriving sequence-structure-function relationships. Experimental characterization of enzymes and designed variants makes heterogeneous data available, though

only as tables in laboratory journals. In an optimal case, these data comprise information on the full sequence of the respective biocatalyst, the measured catalytic activity or selectivity, and the reaction conditions. For publication, the authors usually select a subset of data, which results in a loss of information, e.g. with respect to inactive enzyme variants for a tested set of substrates and reaction conditions. Moreover, the heterogeneous data is distributed in separate tables or figures, which hides the relationships between the applied biocatalyst, the reaction conditions, and the observed results. Despite this drawback, full-text scientific publications including tables and figures are the basis of every study on sequence-function relationships, and a scientist has to find, collect, and integrate this fragmented data into an appropriate data model. To support this process, databases on biochemical properties such as BRENDA (Schomburg et al. 2013), ExplorEnz (McDonald et al. 2007), and KEGG REACTION (Kanehisa et al. 2006) have become available, which incorporate biochemical information obtained by literature mining and provide comprehensive repositories of enzyme function and catalyzed reactions. However, these databases also suffer from the loss of information during the publication process. The lack of the complete protein sequence of a specific biocatalyst or of details on the reaction conditions hinders computational large-scale analysis of sequence-function relationships and, from an experimenter's view, also hinders reproducibility of published experiments.

To fill this gap, comprehensive databases have been suggested that provide comprehensive information on biocatalytic experiments (Apweiler et al. 2010). Instead of parsing biochemical information from published literature, the *BioCatNet* system presented here was designed to collect biochemical information during the experiment and to store it in a consistent and comprehensive data model. Although only a subset of all collected data might be published in a later paper, the complete data is made accessible for future use. Therefore, the *BioCatNet* system can be used as an electronic laboratory journal, which archives heterogeneous experimental data and makes it accessible to systematic searches by the owners of the data or their collaborators. As a first family-specific protein database, the ThDP-dependent Enzyme Engineering Database (TEED) (Widmann, Radloff, and Pleiss 2010) was migrated to *BioCatNet*. ThDP-dependent enzymes form a versatile protein family and catalyze the formation and cleavage of C-C bonds at high chemo- and stereoselectivity. Using the TEED, functionally relevant positions were identified and validated by designing variants with altered activity (Payongsri et al. 2012), as well as chemo- (Yep and McLeish 2009; Yep, Kenyon, and McLeish 2006; Westphal et al. 2014a; Galman et al. 2010; Andrews and McLeish 2012) and stereoselectivity (Westphal et al. 2014a;

Westphal, Jansen, et al. 2014; Westphal, Hahn, et al. 2013; Gocke, Walter, et al. 2008; Rother et al. 2011), thus enabling rational engineering of enzyme variants with desired non-physiological product-forming capabilities. To enable re-interpretation of previous experimental results in a different context and to support rational engineering, biochemical data on ThDP-dependent enzymes will henceforth be systematically related to sequence and structure data.

4.4.3 Concept, Implementation and Population

Linkage of sequence and function

The amino acid sequence of an enzyme is a distinctive trait that can be used to unambiguously identify biocatalysts. Although post-translational processing can result in organism-specific modifications of protein sequences, highly similar sequences commonly fold into similar three-dimensional structures (Browne et al. 1969) and usually resemble each other in their catalytic function. However, single mutations might also cause structural changes (Alexander et al. 2009) or affect the activity and specificity, even if the mutated positions lie outside the active site (Westphal, Jansen, et al. 2014; Drawz et al. 2009; Lingen, Kolter-Jung, et al. 2003). Thus, in order to enable an investigation of the detailed relationships between sequences, structures, and functions of a protein family, all experimental results have to be unambiguously assigned to the exact amino acid sequence of the respective biocatalyst. Specifically, the effect of protein purification tags on activity, stability, and selectivity of an enzyme are often neglected in scientific publications (Majorek et al. 2014). Moreover, a tag might include an additional linker. Therefore, in many publications, the respective biocatalyst is not unambiguously defined, but merely specified by its enzymatic function and its source organism, such as 'pyruvate decarboxylase from *Acetobacter pasteurianus*'. A search for this term in the NCBI protein database demonstrates the ambiguity of such a specification, as it matches 14 protein entries, which differ not only in the length of their N- or C-termini (including truncations and additional His-tags) but also in a total of 130 positions distributed over the entire sequence (Supporting Information, Section A.7.1 on page 225).

Minimal requirements for documentation of biocatalytic experiments

In addition, the biocatalytic properties of an enzyme depend on the reaction conditions, such as pH, temperature, solvent additives, and substrate concentration (Triantafyllou, Adlercreutz, and Mattiasson 1993; Wescott and Klivanov 1994; Gehards et al. 2012; Baraibar et al. 2014). Consequently, for a systematic comparison of biochemical data, detailed information on the respective biocatalyst as well as on the reaction conditions must be given. Therefore, minimal requirements for reporting on biocatalytic results have been suggested by the STRENDA initiative (www.strenda.org) and the European Federation of Biotechnology Section on Applied Biocatalysis (ESAB) (Gardossi et al. 2010). The minimal requirements for data submission to *BioCatNet* were derived from these suggestions in order to find a compromise between the necessary standardization and the usability under laboratory conditions. While only a small amount of information was declared as indispensable for describing biocatalytic experiments (Table 4.6 on the following page), enrichment by additional information is encouraged (Table 4.7 on the next page).

Relational data model

A relational data model adapted to those requirements was implemented in the open source relational database management system *Firebird* (Figure 4.12 on page 137). The tables storing information on protein sequences, proteins, and the family classification were designed on the basis of the respective tables from the *DWARF* data model (Fischer, Thai, et al. 2006) and extended to meet the additional requirements. The additional tables were designed to enable storage of biochemical data and taxonomic lineage. The names of the database tables are given in bold characters in the following.

■ **Amino acid sequence** One of the most eminent entities in the *BioCatNet* data model is the sequence. For each individual amino acid sequence, an entry is defined (**sequences**) that links to a hierarchical classification, the actual amino acid sequence, to information on the respective source of the sequence entries, and to experiments performed with those biocatalysts. Each sequence entry is uniquely defined by the primary key 'sequence_id' (SID). As inherited from the *DWARF* data model, sequences are deposited as a sequence of individual positions (**positions**). Each position of a sequence carries a specific amino acid. In addition,

Table 4.6: Minimal requirements for the submission of biochemical data to *BioCatNet*.

Topic	Mandatory requirements
Scale	Initial volume
Reaction	Reaction description Involved compounds (Educts and Products) Stoichiometry
Biocatalyst(s)	Exact amino acid sequence(s) Sequence name Protein name Source organism Expression host / Host organism in whole-cell biocatalysis Time-point(s) of addition Amount(s) of enzyme Volume(s) of enzyme solution Preparation / Purification / Vendor
Educts	Compound(s) Time-point(s) of addition Amount(s) Volume(s) of additional solvents Preparation / Purification / Vendor
Products	Compound(s) Time-point(s) of measurement Concentration(s) Measurement method
Buffer/Solvent	Unique name, detailed description and concentrations of all components

Table 4.7: Optional information accepted for submission to *BioCatNet*.

Topic	Optional requirements
Additives	Compound(s) Time-point(s) of addition Amount(s) Volume(s) of additional solvents Preparation / Purification / Vendor
Kinetics	k_M k_i $S_{0.5}$ k_{cat} v_{max}
Conditions	pH Temperature Pressure Shaking frequency Further descriptions

standard position numbers as defined by standard numbering schemes can optionally be assigned (Gricman, Vogel, and Pleiss 2014; Vogel, Widmann, et al. 2012; Vogel and Pleiss 2014). In order to allow coloring of sequences and entire multisequence alignments by amino acid properties, or to enable straightforward analysis of conserved physiochemical properties, the amino acids assigned to the individual sequence positions are further specified concerning their names (**aminoacids**) and properties (**aa_property_entries**, **aa_properties**). The position-specific deposition of amino acid sequences further allows annotation of single positions or regions with additional information, such as the experimentally determined functional role of an amino acid (**annotation_entries**, **annotations**).

■ **Entry sources** Each sequence entry is linked to its respective sources; the source organism and the respective online repository, if acquired by an automated homology search. The source entities (**sources**) define unique combinations of sequence entries and source organisms that refer to the respective entries in online repositories (**db_entries**, **db**).

■ **Protein structures and models** *BioCatNet* is capable of linking structure information derived from crystal structures and homology models to the respective amino acid sequences. Structures found in the Protein Data Bank (PDB) (Berman et al. 2000) (**pdb_entries**) or homology models (**model_monomers**) can be assigned to source entries, which implicitly means an assignment to a specific amino acid sequence and the respective source organism. The three-dimensional atom coordinates of crystal structures from the PDB are not explicitly incorporated in the database, but the structures are addressed via the PDB accession code ('ac'), while coordinates of homology models are deposited as flat files and linked by using unique file names ('mo_name'). Since some protein structures in the PDB do not represent the active form of an enzyme (such as apo enzymes without cofactors or non-functional single subunits of multimeric enzymes) and consequently are not suited to serve as templates for automated homology modeling, the data model allows independent deposition of modified template structures (**model_templates**). In case of the Thiamine-diphosphate dependent enzymes Engineering Database (TEED), all structures of family members found in the PDB were modified to construct the active homodimers or heterotetramers. The resulting templates were superimposed and deposited in the TEED. Furthermore, by combination of single monomers generated by homology modeling, the putative structures of multimeric enzymes were constructed based on the structural

information of the respective templates (**multimers**, **model_multimers**).

■ **Hierarchical classification** Sequence entries are embedded into a hierarchical family classification. As established with the *DWARF* system (Fischer, Thai, et al. 2006), sequences belong to proteins (**proteins**), proteins are grouped into homologous families (**h_fam**) based on their global sequence similarity, and homologous families are further grouped into superfamilies (**s_fam**). In *BioCatNet* two additional organizational levels were included to group homologous families and superfamilies according to user-defined criteria (**h_fam_groups**, **s_fam_groups**). In case of the TEED, the groups represent sequence similarity and shared functional annotations, such as multiple homologous families containing 'pyruvate decarboxylases'. Incorporation of the EC numbers (**ec**) further enables grouping of enzymes.

■ **Taxonomy** In order to allow for a taxonomic analysis, the *BioCatNet* system incorporates a copy of the entire taxonomic database as provided by the National Center for Biotechnology Information (NCBI) (Benson et al. 2011) into each family-specific protein database (FSPD). The lineage is represented as a tree starting from the 'root' to the single species and strains (**tax_nodes**). Each node in the lineage is described by a common organism name (**tax_names**) and known synonyms (**tax_synonyms**). Incorporation of the information on merged (**tax_merged**) and deleted (**tax_deleted**) taxonomic identifiers ('tax_id') at the time of the database generation allows comparison between different FSPDs, even if the taxonomic lineage of the NCBI has been modified.

■ **Experiments** One of the major innovations of the *BioCatNet* system is the link between protein sequences and experimental data. Experimental information is organized in experiments (**experiments**) and sets of experiments (**experiment_sets**). Use of experiment sets allows to group different experiments according to user-defined criteria, in order to facilitate handling of the submitted data and to enable straightforward analyses of related results, such as enzyme kinetics or temperature dependencies measured under similar but different reaction conditions. The single experiment links the reaction, the biocatalyst, the reaction conditions and the results. The enzymes used as biocatalysts are defined by the table **enzyme_feeds**. Using this table allows to define multiple biocatalysts to be applied in the same reaction as well as multiple events of enzyme addition. Thus, it supports documentation of fed-batch strategies as well as

cascade reactions. Furthermore, the experiments entity allows the user to control their publication ('release_date'). For each biocatalyst, preparation methods or the respective vendor must be specified (**enzyme_preparation_methods**).

■ **Reactions** *BioCatNet* allows documentation of several reactions simultaneously, if observed within a single experiment (**exp_rkt_link**, **reactions**). Reactions are defined by the educts and the resulting products (**reaction_compounds**). Moreover, different reaction types can be assigned to reactions (**reaction_types**, **rct_type_link**) in order to enable subsequent searches and analyses of different enzymes that catalyze the same reaction types, as well as different reactions and reaction types catalyzed by single enzymes.

■ **Reaction conditions** pH value, temperature, pressure and the shaking frequency are optional parameters that are encouraged to be defined to document the reaction conditions (**conditions**). In addition, the used buffers, including the concentrations of individual buffer components (**buffers**), and the reaction scale are obligatory to be defined.

■ **Compounds: substrates, products and additives** All compounds, regardless of whether later defined as educts, products, or additives in any documented reaction, are stored by the respective isomeric SMILES codes (**compounds**). Compound names and SMILES codes can be requested from the PubChem database (Bolton et al. 2008) or added by the user (**compound_names**). For both, substrates and additives, the added amount and the respective time-point of addition to the experiment have to be reported (**substrate_feeds**, **additive_feeds**). In case of changing the reaction volume by addition of solved substrates or additives, the volume of the respective solvent has to be documented. Furthermore, the preparation of the compounds or the vendor have to be reported (**compound_preparation_methods**).

■ **Methods and product concentration** The methods of product detection (**methods**) and the amount of measured product (**compound_measurements**) have to be defined. Furthermore, *BioCatNet* allows to document the product concentration as well as the concentrations and the respective detection methods of other compounds measured at various time-points.

Miscellaneous In three further tables, the units are defined (**units**), literature information is assigned to experiments (**literature**, **literature_entries**), and previously removed sequence entries are documented to avoid repeated insertion upon updates of the FSPD (blacklist) (**seen_db_entries**).

Derived attributes

In most publications, the reported results of biocatalytic experiments are parameters that have been derived by fitting experimental data to a kinetic model. Thus, the results from two experiments can only be compared if an identical kinetic model was used. Consequently, it is preferable to also provide access to the raw data and allow users to analyze them by applying different kinetic models. A similar strategy is followed in depositing protein structures in the Protein Data Bank. In addition to the result of structure refinement (the atomic coordinates and B-factors), the experimental data are deposited to allow for alternative refinements. According to the same principles, enantiomeric excesses are likewise not recorded in *BioCatNet* but represented by the defined amounts of (*S*)- or (*R*)-enantiomers of educts and products used or yielded, respectively.

User management

To protect proprietary information, *BioCatNet* manages permissions of users and groups. Newly inserted amino acid sequences and experiments are owned by specific users. Only the owners are allowed to retrieve, edit or delete the data sets, or to release them to the public.

Parsing and population

BioCatNet combines two different strategies of data acquisition:

1. Sequences and structures of enzymes are automatically collected from public repositories and incorporated into the respective FSPD, while
2. biochemical information is requested from bench scientists (Figure 4.13 on page 138).

The latter is, as discussed later, facilitated by a web-accessible graphical user interface, whereas the initial collection of sequence and structure information on a specific protein family ('parsing')

and the subsequent insertion into the FSPD ('pushing') is done using a toolbox of Perl scripts (*DBParse*).

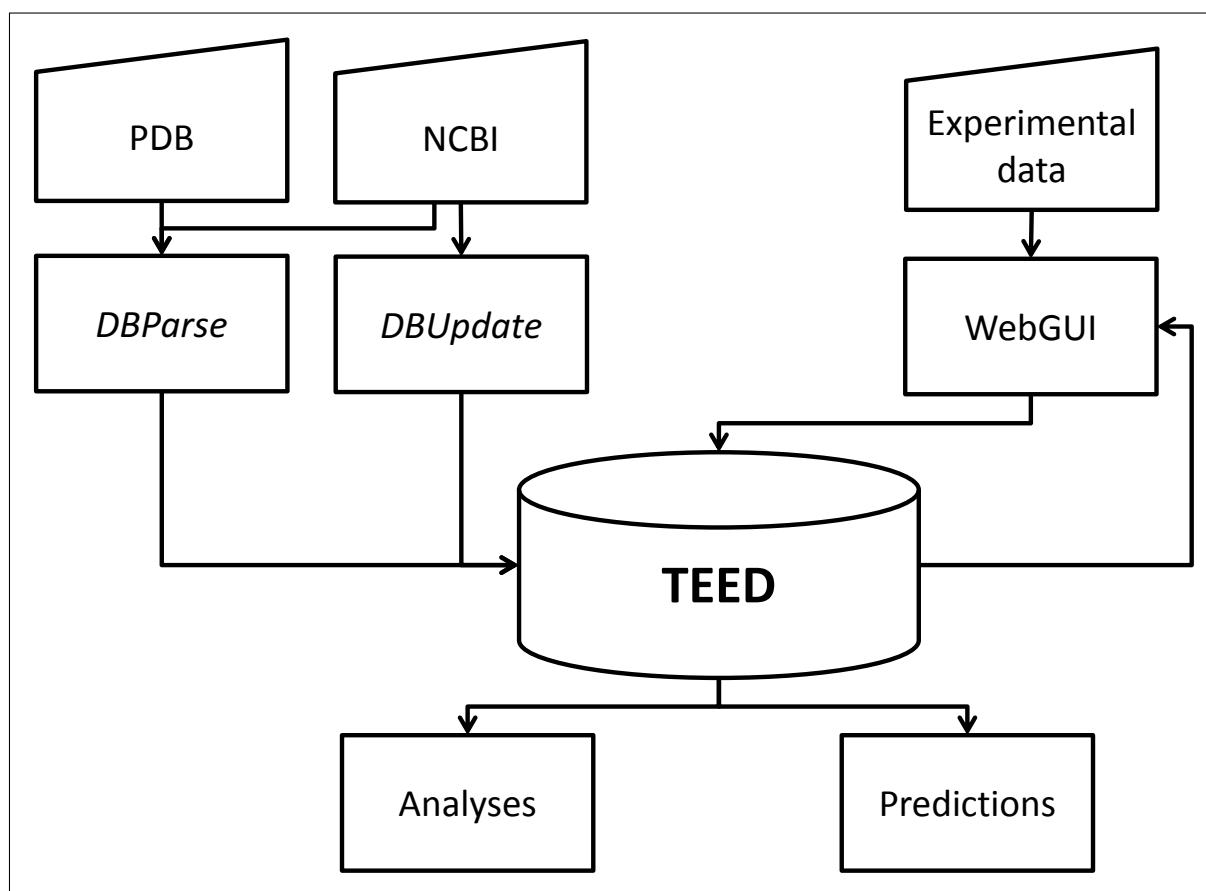


Figure 4.13: Schematic representation of the *BioCatNet* system. The *BioCatNet* system is able to generate and maintain family-specific protein databases, as exemplary shown for the TEED. Structure and sequence information is acquired by the tools *DBParse* and *DBUpdate*, which extract information about proteins belonging to defined families from the Protein DataBank (Berman et al. 2000) and the non-redundant protein database provided by the NCBI Protein database (Benson et al. 2011). Input of biochemical data as well as visual inspection, editing and manual analyses are enabled via a web-accessible graphical user interface (WebGUI). The data on sequence, structure and function of enzymes can subsequently be used for systematic analyses of the sequence-structure-function relationships and the prediction of promising variants for rational engineering.

A protein family is defined by a set of representatives, the seed sequences, which are in subsequent steps used to identify homologous proteins and to sort the identified sequences into homologous families. Therefore, the seed sequences serve as queries of *blastp* (Camacho et al. 2009) searches in the non-redundant protein repository provided by the NCBI (Benson et al. 2011) with a user-defined expectation threshold. The resulting accession numbers of homologous sequences are subsequently used to download the complete entries in XML format. The amino acid sequences of all downloaded entries are extracted from the XML files and used as the basis for the family classification:

1. Identical sequences (including fragments) are grouped to sequence entries using *usearch* (Camacho et al. 2009) and only the longest sequence is used for further classification (sequence representative).
2. Sequence entries with identities of more than 98% between their sequences representatives are grouped into protein entries and the longest sequence of each protein entry is used for further classification (protein representative).
3. Subsequently, the set of seed sequences is used for a first classification into homologous families. Therefore, all protein representatives are compared to all seed sequences using *ublast* (Camacho et al. 2009) to identify the nearest seed sequence of each protein. The protein representatives are then globally aligned to their most similar seed sequence by *needle* from the EMBOSS package (Rice, Longden, and Bleasby 2000). If the similarity of both aligned sequences is higher than 40%, the protein entry is assigned to the homologous family defined by the respective seed sequence.
4. Proteins, which are less similar to the seed sequences, are subsequently compared to all sequences that have been assigned to homologous families in the previous step. Sequences with global sequence similarities higher than 40% to any of the already assigned sequences are grouped to the respective homologous families. Less similar sequences are sorted into additional homologous families.

The results from the family classification are subsequently used to create the FSPD. Therefore, the pre-calculated homologous families, protein and sequence entries are pushed into the database. For each sequence entry, accession numbers, the protein name, annotations, the taxon ID and, if defined, the EC number are parsed from the XML files and pushed into the FSPD.

Structure information is subsequently derived from the Protein Data Bank (PDB) (Berman et al. 2000). Therefore, all sequences of structures in the PDB are downloaded and compared to the sequences that were previously included in the FSPD. Structures homologous to sequences in the FSPD are classified to sequence and protein entries using the same criteria as mentioned above.

Furthermore, an update routine (*DBUpdate*) was implemented, allowing for regular actualization of the sequence and structure information stored in the FSPD. Therefore, the sequences in the FSPD are clustered using *usearch* (Camacho et al. 2009) with an identity threshold of 30% and the resulting centroids are used as representatives for homology searches in the NCBI protein database. The XML files of newly identified sequences that are not listed on the blacklist

(table: **seen_db_entries**) are downloaded and the sequences are included into the FSPD without changing the family classification.

Curation

Generation and maintenance of a FSPD requires prior knowledge about the respective protein family. Although most steps during the initial generation and subsequent updates of the FSPD were automated with *DBParse* and *DBUpdate*, several steps require intervention of a curator, who adds biochemical knowledge such as the initial selection of seed sequences, the choice of an expectation threshold for the search for homologous sequences, the classification of homologous families into superfamilies, further grouping of homologous families and superfamilies, as well as manual revision of the automatically inserted homologous families.

Architecture

The *BioCatNet* is running on a Debian 4.6.3 system, using Perl 5.14, PHP 5.4.4, Apache 2.2.22 and Firebird 2.5. The parsing and pushing algorithms were implemented in Perl and optimized for multithreading. The back- and frontend of the graphical user interface were developed in PHP. Due to the applied client-server-architecture, the *BioCatNet* WebGUI is accessible using any internet browser. In order to enable confidential data transfer during submission of experimental results, all traffic is secured by TLS/SSL.

Graphical user interface

The *BioCatNet* web interface provides specialized views for browsing the sequence or structure space, analyzing the organism distribution, exploring the catalytic potential of the respective enzyme family, or adding information.

The *Sequences* view provides access through a hierarchy organized by superfamily-groups, superfamilies, homologous family-groups, homologous families, proteins and sequence entries. It is highly enriched with additional information (Table 4.8 on page 142). For all superfamilies and homologous families, the sequences can be downloaded in FASTA format, as well as multisequence alignments and phylogenetic trees generated by *clustalo* (Sievers et al. 2011). By using the sequence alignment viewer *JalView* (Waterhouse et al. 2009), the multisequence

alignments and phylogenetic trees can be dynamically visualized. On the lowest level of the hierarchically organized Sequence view, each sequence and all information about this sequence can be viewed (Figure 4.14 on page 145); annotation of the functional role of individual residues, assigned standard numbering schemes (Gricman, Vogel, and Pleiss 2014; Vogel, Widmann, et al. 2012), the source organisms, documented biochemical functions, function information inferred from EC numbers, structure information from the PDB or homology models, and links to further repositories on sequence, structure, and function. Implementation of the *PV* protein viewer (<https://biasmv.github.io/pv>) enables interactive inspection of pre-generated homology models and superimpositions with the respective template structures in WebGL compatible browsers.

The *Structures* view summarizes the available crystal structure entries sorted by the hierarchy mentioned above.

The *Taxonomy* view allows for browsing through the lineage of organisms. Thereby it summarizes, for which taxonomic nodes or subordinates, sequence, structure or function information is available.

The *Functions* view provides access to all released experiments that have been entered into the FSPD. Since the *BioCatNet* system is going to be released for public usage simultaneously with publication of this article, the *Functions* view does not show experimental results, yet.

Data input is accomplished via the *Workbench*, which is exclusively accessible for authorized users. In order to facilitate data submission, the *Workbench* summarizes all previously added experiment sets and experiments of the respective user, as well as sequences of the respective enzymes. Moreover, implementation of a caching engine enables discontinuous data input without the risk of losing previously entered information. Implementation of the chemical compound drawing tool *Ketcher* from GGA Software Services (Karulin and Kozhevnikov 2011) further enables an intuitive definition of substrates, additives, and products. In order to minimize the effort of data submission, the *Workbench* interactively supports data input with an auto-complete capability and provides the experimenter with a formatted printout summarizing all submitted data, which can be used in laboratory journals or scientific reports.

Table 4.8: Additional information available for the different hierarchical levels in the *Sequence* view of the TEED.

Information	Level ^[a]
Distribution of homologous families	DB, SF
Distribution of protein entries	DB, SF, HF
Distribution of sequence entries	DB, SF, HF, P
Distribution of structures	DB, SF, HF, P, S
Organism distribution	DB, SF, HP, P, S
Averaged sequence length	HF
Distribution of sequence lengths	P
Information on the structure and sequence architecture	SF
Annotated multisequence alignments	SF, HF
Download of all sequences	DB, SF, HF, P, S
Distribution of EC numbers	HF

^[a] DB, entire database; SF, superfamily; HF, homologous family; P, protein; S, sequence

4.4.4 Application and future perspectives

A FSPD on ThDP-dependent enzymes, the TEED (www.teed.biocatnet.de), was recently updated and migrated to the *BioCatNet* system (Vogel and Pleiss 2014). 51 representative sequences of ThDP-dependent enzymes were used as seed sequences and searched against NCBI (Benson et al. 2011) and PDB (Berman et al. 2000) for homologous sequences and structures (Supporting Information, Table A.5 on page 210). 77493 unique, homologous sequences were collected, assigned to 52565 different protein entries, and classified into 168 homologous families and in nine superfamilies (Supporting Information, Table A.9 on page 228). The superfamilies were defined based on functional annotation and structural architecture (Vogel and Pleiss 2014). 240 crystal structures from the PDB were linked to 284 respective sequence entries. All sequences in the TEED were checked for similarity to one of the 240 structures. For a global sequence similarity of at least 60%, automatic homology modeling was performed by MODELLER 9.13 (Fiser and Sali 2003; Webb and Sali 2014). Thus, the fraction of thiamine diphosphate-dependent enzymes with available structure information was increased from 0.4 to 31%.

Similarly, family-specific protein databases on imine reductases (Scheller et al. 2014), cytochrome P450 monooxygenases (Nelson 2009; Fischer, Knoll, et al. 2007; Gricman, Vogel, and Pleiss 2014; Sirim, Wagner, Lisitsa, et al. 2009; Sirim, Widmann, et al. 2010), TEM lactamases, and triterpene cyclases (Racolta et al. 2012) are currently migrated to *BioCatNet* to serve as valuable information sources on sequence, structure and biochemical data. The

system aids in efficiently documenting heterogeneous experimental data by providing a comprehensive data model. The data model of *BioCatNet* defines minimal requirements for the documentation of enzyme-catalyzed reactions, and thus follows the recommendations of the STRENDA guidelines for reporting biocatalytic reactions (Gardossi et al. 2010) (www.beilstein-institut.de/en/projects/strenda).

Because all biochemical data is entered manually by experimenters, *BioCatNet* prevents the drawbacks of reconstructing heterogeneous information from literature, where biocatalytic data on an enzyme, its variants, and catalyzed reactions is scattered over different sections, figures, tables and supplementary materials of multiple publications. Moreover, in the process from planning and performing biocatalytic experiments towards the publication of results, findings which are less relevant for the focus of the manuscript are lost. However, in the future, these primary results might gain relevance, though in a changed context due to a deeper knowledge about the family-specific sequence-structure-function relationships. *BioCatNet* is intended to access the invaluable wealth of information about biochemical data available in experienced experimental laboratories to enable future analyses of the primary experimental data in different contexts, or to compare it to data generated by collaborators on different homologous enzymes. In contrast to the conventional strategies of data management, all experimental data (whether published or not) is made accessible, either for exchange with collaborators or for a future analysis when the original data might be re-inspected with a different scientific perspective.

In order to allow for comparison between homologous enzymes, *BioCatNet* implements family-specific standard numbering schemes that enable an identification of corresponding positions (Vogel, Widmann, et al. 2012; Vogel and Pleiss 2014; Gricman, Vogel, and Pleiss 2014). By combining a standard numbering scheme with information about the functional roles of individual positions in different enzymes, *BioCatNet* supports prediction of functionally relevant positions. In different ThDP-dependent enzymes, standard position 477 has been identified to affect the stereoselectivity. According to the *S*-pocket concept (Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008; Rother et al. 2011), a bulky side chain at standard position 477 endorses formation of (*R*)-product, whereas small side chains shift the stereoselectivity towards the (*S*)-product (Sergienko and Jordan 2001a; Lingen, Kolter-Jung, et al. 2003; Gocke, Walter, et al. 2008; Meyer, Walter, et al. 2011; Rother et al. 2011; Westphal, Hahn, et al. 2013; Westphal, Waltzer, et al. 2013; Westphal et al. 2014a; Westphal, Jansen, et al. 2014). By identification of the corresponding position in an enzyme of interest, *BioCatNet* is able to predict a promising mutation target for

the generation of variants with altered stereoselectivity.

With this knowledge on the role of standard position 477, previous unpublished data was re-inspected. In 2006, the gene encoding a potential benzaldehyde lyase (BAL) from *Rhodospirillum rubrum* (*RpBAL*, NCBI genbank entry RPA0108, gi|39933188) was cloned, extended by a C-terminal hexahistidine residue, expressed in *E. coli* BL21(DE3) and successfully purified (Brosi 2006). In contrast to the well-described, (*R*)-specific BAL from *Pseudomonas fluorescens* (*PfBAL*, gi|9965498, pdb|2AG0) (Demir, Sesenoglu, Eren, et al. 2002; Demir, Sesenoglu, Dünkelmann, et al. 2003; Domínguez de María et al. 2007; Mosbacher, Müller, and Schulz 2005; Brandt, Kneen, et al. 2010; Brandt, Nemeria, et al. 2008), *RpBAL* did not show activity towards benzoin but weak carbonylation activity forming (*S*)-acetoin from acetaldehyde with 60% ee. Due to its low activity, it was not considered as a useful member of the toolbox of ThDP-dependent enzymes for the synthesis of chiral 2-hydroxy ketones (Pohl, Gocke, and Müller 2010) and was therefore not further characterized. However, the *S*-pocket concept postulated later (Knoll, Müller, et al. 2006; Gocke, Walter, et al. 2008; Rother et al. 2011) now puts a different complexion on *RpBAL*. At the three standard positions 28, 477 and 480 lining the *S*-pocket (Rother et al. 2011), *RpBAL* possesses small side chains (two glycines and a serine), in contrast to *PfBAL* (alanine, phenylalanine and threonine), thus resulting in a wider *S*-pocket as compared to *PfBAL*. The experiments performed in 2006, though performed with another intention, nicely approve the later postulated *S*-pocket concept.

Consequently, deposition of successful and seemingly less successful experiments will not only reduce repeated attempts to convert the same substrates or substrates not accepted by certain enzymes, but it also offers the opportunity to re-interpret data as well as to design new experiments in the light of previous attempts. It also offers the opportunity to detect links between published or unpublished data and to re-interpret previous observations to gain a more complete understanding of the sequence-structure-function relationships of an enzyme family. As a consequence, rational design of optimized biocatalysts will benefit from the gain of biochemical information and the possibility to transfer knowledge between different enzymes.

all superfamilies / [SF#1] DC-like / [SF#1] decarboxylases (DC) / [HF#2] PDC-like / [HF#3] IPDC / [P#1543] indole-3-pyruvate decarboxylase / [S#2854] indole-3-pyruvate decarboxylase

[S#2854] indole-3-pyruvate decarboxylase

([P#1543] indole-3-pyruvate decarboxylase)

print overview
copy FASTA to clipboard
download FASTA
ExplorEnz
KEGG
BRENDA
standard positions

Sequence

>sid|2854|pid|1543|hfid|3|sfid|1|emb|CBX71533.1|taxid|913028|

```
MASNYKVADY LLDRLAQVGI RHLFGVPGDF NLHFLDHVIS HPVIQWMGCA NELNAAAYAD 60
GYARVMPAAA LLTTTGVGEL SAINGIAGSF AEYLPITIHV GTPALRSQKA GELLHHSFGD 120
GDFNHFARMA KEVACAHTSL TAENAASEID RLLVAALYQR RPVYLQLP SD VGEAELTSQS 180
GVLALSQPML SPTSLQAFIE AARQKLQSAI -----ARQALN HWLAEVNLPH 240
STLLMGKGLL DETHPMFIGT YAGAASDAS I 517 (527) VWFVDT ITAGFSQHIT 300
QDNCIDVQPE QVRIGRQVFS QIPMLAAVN -----PVI AHS MPALPCDNLL 360
SQQALWYHIQ HFLRPDDIV TDQGTSSFQ PP (363..528) SLWGSI GFSLPAAYGA 420
QLAQPPRRVI LLVGDGAAQL TIQELGSMLR DGLTF -----LLINNDYTVERA IHGPQQPYND 480
IAEWDWTQLP QALSVDKASL TCRVTQADEL QQVLIQIENC QQLAFIEV HL PPMDMPELMI 540
NVAKSIQARN AAV
[anno#141] PYR
[anno#142] PP
```

Sources

[T#913028] Yersinia enterocolitica W22703 (?) (ncbi)	emb CBX71533.1 gi 330861288
[T#994476] Yersinia enterocolitica subsp. palearctica 105.5R(r) (?) (ncbi)	gi 332162512

Inferred Structures (by homology modelling)

model	template	GA341
[HM#22428]	pdb 10VM (chains: A, B, C, D) [S#2715] RecName: Full=Indole-3-pyruvate decarboxylase; Short=Indolepyruvate decarboxylase [T#550] Enterobacter cloacae (?) (ncbi)	1.000000 ↓

Documented Functions

No experiments with this protein sequence have been posted yet

Inferred functions





Source	Educt	Product	Pathways
 rn:R00014	Pyruvate + Thiamin diphosphate	2-(alpha-Hydroxyethyl)thiamine diphosphate + CO ₂	Glycolysis / Gluconeogenesis Citrate cycle (TCA cycle) Pyruvate metabolism Biosynthesis of secondary metabolites Microbial metabolism in diverse environments
 rn:R00224	Pyruvate	Acetaldehyde + CO ₂	
 rn:R00636	2-Oxo acid	Aldehyde + CO ₂	
 rn:R00755	Acetaldehyde + Thiamin diphosphate	2-(alpha-Hydroxyethyl)thiamine diphosphate	Glycolysis / Gluconeogenesis Metabolic pathways

Figure 4.14: Sequence view of the BioCatNet WebGUI. Each sequence of a family-specific protein database is linked to information on eventually existing crystal structures or homology models, and is enriched by annotations. Furthermore, it provides access to the respective sources in online repositories on sequences, structures, functions and taxonomy of enzymes. If available, function information submitted by experimenters or inferred from an assigned EC number is provided.

References

- Abhinandan, KR and Martin, ACR (2008). Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol Immunol* **45** (14), 3832–3839.
- Alexander, PA, He, Y, Chen, Y, Orban, J, and Bryan, PN (2009). A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci USA* **106.50**, 21149–21154.
- Altschul, SF, Gish, W, Miller, W, Myers, EW, and Lipman, DJ (1990). Basic local alignment search tool. *J Mol Biol* **215** (3), 403–413.
- Ambler, RP (1980). The structure of beta-lactamases. *Philos Trans R Soc Lond B Biol Sci* **289**, 321–331.
- Ambler, RP, Coulson, AF, Frère, JM, Ghuysen, JM, Joris, B, Forsman, M, Levesque, RC, Tiraby, G, and Waley, SG (1991). A standard numbering scheme for the class A beta-lactamases. *Biochem J* **276**, 269–270.
- Anbar, M, Gul, O, Lamed, R, Sezerman, UO, and Bayer, EA (2012). Improved Thermostability of *Clostridium thermocellum* Endoglucanase Cel8A by Using Consensus-Guided Mutagenesis. *Appl Environ Microbiol* **78.9**, 3458–3464.
- Andrews, FH and McLeish, MJ (2012). Substrate specificity in thiamin diphosphate-dependent decarboxylases. *Bioorg Chem* **43**, 26–36.
- Andrews, FH, Rogers, MP, Paul, LN, and McLeish, MJ (2014). Perturbation of the Monomer-Monomer Interfaces of the Benzoylformate Decarboxylase Tetramer. *Biochemistry* **53** (27), 4358–4367.
- Anfinsen, CB (1973). Principles that Govern the Folding of Protein Chains. *Science* **181** (4096), 223–230.
- Apweiler, R, Armstrong, R, Bairoch, A, Cornish-Bowden, A, Halling, PJ, Hofmeyr, JS, Kettner, C, Leyh, TS, Rohwer, J, Schomburg, D, Steinbeck, C, and Tipton, K (2010). A large-scale protein-function database. *Nat Chem Biol* **6**, 785.

- Arjunan, P, Sax, M, Brunskill, A, Chandrasekhar, K, Nemeria, N, Zhang, S, Jordan, F, and Furey, W (2006). A Thiamin-bound, Pre-decarboxylation Reaction Intermediate Analogue in the Pyruvate Dehydrogenase E1 Subunit Induces Large Scale Disorder-to-Order Transformations in the Enzyme and Reveals Novel Structural Features in the Covalently Bound Adduct. *J Biol Chem* **281** (22), 15296–15303.
- Arnold, FH (2001). Combinatorial and computational challenges for biocatalyst design. *Nature* **409**, 253–257.
- Arnold, K, Bordoli, L, Kopp, J, and Schwede, T (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22** (2), 195–201.
- Asztalos, P, Parthier, C, Golbik, R, Kleinschmidt, M, Hübner, G, Weiss, MS, Friedemann, R, Wille, G, and Tittmann, K (2007). Strain and Near Attack Conformers in Enzymic Thiamin Catalysis: X-ray Crystallographic Snapshots of Bacterial Transketolase in Covalent Complex with Donor Ketoses Xylulose 5-phosphate and Fructose 6-phosphate, and in Noncovalent Complex with Acceptor Aldose Ribose 5-phosphate. *Biochemistry* **46** (43), 12037–12052.
- Baburina, I, Dikdan, G, Guo, F, Tous, GI, Root, B, and Jordan, F (1998). Reactivity at the substrate activation site of yeast pyruvate decarboxylase: inhibition by distortion of domain interactions. *Biochemistry* **37**, 1245–1255.
- Baburina, I, Gao, Y, Hu, Z, Jordan, F, Hohmann, S, and Furey, W (1994). Substrate activation of brewers' yeast pyruvate decarboxylase is abolished by mutation of cysteine 221 to serine. *Biochemistry* **33**, 5630–5635.
- Baburina, I, Li, H, Bennion, B, Furey, W, and Jordan, F (1998). Interdomain Information Transfer during Substrate Activation of Yeast Pyruvate Decarboxylase: The Interaction between Cysteine 221 and Histidine 92. *Biochemistry* **37** (5), 1235–1244.
- Baig, IA, Gedi, V, Lee, SC, Koh, SH, and Yoon, MY (2013). Role of a highly conserved proline-126 in ThDP binding of *Mycobacterium tuberculosis* acetohydroxyacid synthase. *Enzyme Microb Technol* **53** (4), 243–249.
- Baker, D (2000). A surprising simplicity to protein folding. *Nature* **405**, 39–42.
- Balakrishnan, A, Gao, Y, Moorjani, P, Nemeria, NS, Tittmann, K, and Jordan, F (2012). Bifunctionality of the Thiamin Diphosphate Cofactor: Assignment of Tautomeric/Ionization States of the 4'-Aminopyrimidine Ring When Various Intermediates Occupy the Active Sites during the Catalysis of Yeast Pyruvate Decarboxylase. *J Am Chem Soc* **134** (8), 3873–3885.

- Balakrishnan, A, Nemeria, NS, Chakraborty, S, Kakalis, L, and Jordan, F (2012). Determination of Pre-Steady-State Rate Constants on the *Escherichia coli* Pyruvate Dehydrogenase Complex Reveals That Loop Movement Controls the Rate-Limiting Step. *J Am Chem Soc* **134** (45), 18644–18655.
- Baraibar, AG, Lieres, E von, Wiechert, W, Pohl, M, and Rother, D (2014). Effective Production of (*S*)- α -Hydroxy ketones: An Reaction Engineering Approach. *Topics in Catalysis* **57.5**, 401–411.
- Bar-Ilan, A, Balan, V, Tittmann, K, Golbik, R, Vyazmensky, M, Hübner, G, Barak, Z, and Chipman, DM (2001). Binding and Activation of Thiamin Diphosphate in Acetohydroxyacid Synthase. *Biochemistry* **40** (39), 11946–11954.
- Barth, S, Fischer, M, Schmid, RD, and Pleiss, J (2004a). Sequence and structure of epoxide hydrolases: A systematic analysis. *Proteins* **55.4**, 846–855.
- (2004b). The database of epoxide hydrolases and haloalkane dehalogenases: one structure, many functions. *Bioinformatics* **20** (16), 2845–2847.
- Baugh, L, Gallagher, LA, Patrapuvich, R, Clifton, MC, Gardberg, AS, Edwards, TE, Armour, B, Begley, DW, Dieterich, SH, Dranow, DM, Abendroth, J, Fairman, JW, Fox, D, Staker, BL, Phan, I, Gillespie, A, Choi, R, Nakazawa-Hewitt, S, Nguyen, MT, Napuli, A, Barrett, L, Buchko, GW, Stacy, R, Myler, PJ, Stewart, LJ, Manoil, C, and Van Voorhis, WC (2013). Combining Functional and Structural Genomics to Sample the Essential *Burkholderia* Structome. *PLoS ONE* **8** (1), e53851.
- Baykal, A, Chakraborty, S, Dodoo, A, and Jordan, F (2006). Synthesis with good enantiomeric excess of both enantiomers of α -ketols and acetolactates by two thiamin diphosphate-dependent decarboxylases. *Bioorg Chem* **34** (6), 380–393.
- Beigi, M, Waltzer, S, Fries, A, Eggeling, L, Sprenger, GA, and Müller, M (2013). TCA Cycle Involved Enzymes SucA and Kgd, as well as MenD: Efficient Biocatalysts for Asymmetric C-C Bond Formation. *Org Lett* **15** (3), 452–455.
- Belenky, I, Steinmetz, A, Vyazmensky, M, Barak, Z, Tittmann, K, and Chipman, DM (2012). Many of the functional differences between acetohydroxyacid synthase (AHAS) isozyme I and other AHASs are a result of the rapid formation and breakdown of the covalent acetolactate-thiamin diphosphate adduct in AHAS-I. *FEBS J* **279** (11), 1967–1979.
- Benson, DA, Karsch-Mizrachi, I, Lipman, DJ, Ostell, J, and Sayers, EW (2011). GenBank. *Nucleic Acids Res* **39**, D32–D37.

- Berendsen, HJC, Grigera, JR, and Straatsma, TP (1987). The missing term in effective pair potentials. *J Chem Phys* **91** (24), 6269–6271.
- Berg, JM, Tymoczko, JL, and Stryer, L (2014). *Biochemie*. 7th ed. Springer-Verlag Berlin Heidelberg.
- Berheide, M, Peper, S, Kara, S, Long, WS, Schenkel, S, Pohl, M, Niemeyer, B, and Liese, A (2010). Influence of the hydrostatic pressure and pH on the asymmetric 2-hydroxyketone formation catalyzed by *Pseudomonas putida* benzoylformate decarboxylase and variants thereof. *Biotechnol Bioeng* **106** (1), 18–26.
- Berman, HM, Westbrook, J, Feng, Z, Gilliland, G, Bhat, TN, Weissig, H, Shindyalov, IN, and Bourne, PE (2000). The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242.
- Berthold, CL, Moussatche, P, Richards, NGJ, and Lindqvist, Y (2005). Structural basis for activation of the thiamin diphosphate-dependent enzyme oxalyl-CoA decarboxylase by adenosine diphosphate. *J Biol Chem* **280**, 41645–41654.
- Berthold, CL, Toyota, CG, Moussatche, P, Wood, MD, Leeper, F, Richards, NGJ, and Lindqvist, Y (2007). Crystallographic Snapshots of Oxalyl-CoA Decarboxylase Give Insights into Catalysis by Nonoxidative ThDP-Dependent Decarboxylases. *Structure* **15** (7), 853–861.
- Bhasin, M, Billinsky, JL, and Palmer, David RJ (2003). Steady-State Kinetics and Molecular Evolution of *Escherichia coli* MenD [(1R,6R)-2-Succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate Synthase], an Anomalous Thiamin Diphosphate-Dependent Decarboxylase-Carboligase. *Biochemistry* **42** (46), 13496–13504.
- Blundell, TL, Sibanda, BL, Sternberg, MJW, and Thornton, JM (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**, 347–352.
- Bobst, CE and Tabita, FR (2004). The role of cysteine 160 in thiamine diphosphate binding of the Calvin-Benson-Bassham cycle transketolase of *Rhodobacter sphaeroides*. *Arch Biochem Biophys* **426** (1), 43–54.
- Bolton, EE, Wang, Y, Thiessen, PA, and Bryant, SH (2008). Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry*. Ed. by RA Wheeler and DC Spellmeyer. **4**. Elsevier, 217–241.
- Bradford, MM (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**, 248–254.

- Brandt, GS, Kneen, MM, Petsko, GA, Ringe, D, and McLeish, MJ (2010). Active-site engineering of benzaldehyde lyase shows that a point mutation can confer both new reactivity and susceptibility to mechanism-based inhibition. *J Am Chem Soc* **132**, 438–439.
- Brandt, GS, Nemeria, N, Chakraborty, S, McLeish, MJ, Yep, A, Kenyon, GL, Petsko, GA, Jordan, F, and Ringe, D (2008). Probing the active center of benzaldehyde lyase with substitutions and the pseudosubstrate analogue benzoylphosphonic acid methyl ester. *Biochemistry* **47**, 7734–7743.
- Breslow, R (1957). Rapid deuterium exchange in tetrazolium salts. *J Am Chem Soc* **79**, 1762.
- Breuer, M, Pohl, M, Hauer, B, and Lingen, B (2002). High-throughput assay of (*R*)-phenylacetylcarbinol synthesized by pyruvate decarboxylase. *Anal Bioanal Chem* **374** (6), 1069–1073.
- Brosi, Helen (2006). Klonierung und Charakterisierung neuer Thiamindiphosphat-abhängiger Enzyme. Diploma thesis. RWTH Aachen, University.
- Brovetto, M, Gamemara, D, Míndez, PS, and Seoane, GA (2011). C-C Bond-Forming Lyases in Organic Synthesis. *Chem Rev* **111** (7), 4346–4403.
- Browne, WJ, North, ACT, Phillips, DC, Brew, K, Vanaman, TC, and Hill, RL (1969). A possible three-dimensional structure of bovine α -lactalbumin based on that of hen's egg-white lysozyme. *J Mol Biol* **42** (1), 65–86.
- Bruhn, H, Pohl, M, Grötzinger, J, and Kula, MR (1995). The replacement of Trp392 by alanine influences the decarboxylase/carboligase activity and stability of pyruvate decarboxylase from *Zymomonas mobilis*. *Eur J Biochem* **234**, 650–655.
- Camacho, C, Coulouris, G, Avagyan, V, Ma, N, Papadopoulos, J, Bealer, K, and Madden, TL (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.
- Candy, JM and Duggleby, RG (1994). Investigation of the cofactor-binding site of *Zymomonas mobilis* pyruvate decarboxylase by site-directed mutagenesis. *Biochem J* **300** (Pt 1), 7–13.
- (1998). Structure and properties of pyruvate decarboxylase and site-directed mutagenesis of the *Zymomonas mobilis* enzyme. *Biochim Biophys Acta* **1385** (2), 323–338.
- Candy, JM, Koga, J, Nixon, PF, and Duggleby, RG (1996). The role of residues glutamate-50 and phenylalanine-496 in *Zymomonas mobilis* pyruvate decarboxylase. *Biochem J* **315**, 745–751.
- Chabrière, E, Charon, MH, Volbeda, A, Pieulle, L, Hatchikian, EC, and Fontecilla-Camps, JC (1999). Crystal structures of the key anaerobic enzyme pyruvate:ferredoxin oxidoreductase, free and in complex with pyruvate. *Nat Struct Biol* **6**, 182–190.

- Chang, AK, Nixon, PF, and Duggleby, RG (1999). Aspartate-27 and glutamate-473 are involved in catalysis by *Zymomonas mobilis* pyruvate decarboxylase. *Biochem J* **339**, 225–260.
- Chong, CK, Shin, HJ, Chang, SI, and Choi, JD (1999). Role of Tryptophanyl Residues in Tobacco Acetolactate Synthase. *Biochem Biophys Res Commun* **259** (1), 136–140.
- Chothia, C (1992). One thousand families for the molecular biologist. *Nature* **357**, 543–544.
- Chothia, C and Lesk, AM (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J* **5** (4), 823–826.
- Chung, ST, Tan, RTY, and Suzuku, I (1971). Glyoxylate carboligase of *Pseudomonas oxalaticus*. Possible structural role for flavine-adenine dinucleotide. *Biochemistry* **10**, 1205–1209.
- Costelloe, SJ, Ward, JM, and Dalby, PA (2008). Evolutionary analysis of the TPP-dependent enzyme family. *J Mol Evol* **66**, 36–49.
- D'Angelo, G and Rampone, S (2014). Towards a HPC-oriented parallel implementation of a learning algorithm for bioinformatics applications. *BMC Bioinformatics* **15** (Suppl 5), S2.
- Dawson, A, Chen, M, Fyfe, PK, Guo, Z, and Hunter, WN (2010). Structure and reactivity of *Bacillus subtilis* MenD catalyzing the first committed step in menaquinone biosynthesis. *J Mol Biol* **401** (2), 253–264.
- Dayhoff, MG, Eck, RV, Chang, MA, and Sochard, MR (1965). Atlas of Protein Sequence and Structure. Silver Spring: National Biomedical Research Foundation, Maryland, USA.
- Demir, AS, Dünwald, T, Iding, H, Pohl, M, and Müller, M (1999). Asymmetric benzoin reaction catalyzed by benzoylformate decarboxylase. *Tetrahedron Asymmetry* **10** (24), 4769–4774.
- Demir, AS, Peruze, A, Umut, D, and Umut, J (2008). *Fusarium roseum* and *Aspergillus oryzae*-mediated enantioselective reduction of benzils to benzoin. *J Mol Catal B Enzym* **55** (3-4), 164–168.
- Demir, AS, Pohl, M, Janzen, E, and Müller, M (2001). Enantioselective synthesis of hydroxy ketones through cleavage and formation of acyloin linkage. Enzymatic kinetic resolution via C-C bond cleavage. *J Chem Soc Perkin I* **1**, 633–635.
- Demir, AS, Sesenoglu, Ö, Dünkelmann, P, and Müller, M (2003). Benzaldehyde lyase-catalyzed enantioselective carbonylation of aromatic aldehydes with mono- and dimethoxy acetaldehyde. *Org Lett* **5**, 2047–2050.
- Demir, AS, Sesenoglu, Ö, Eren, E, Hosrik, B, Pohl, M, Janzen, E, Kolter, D, Feldmann, R, Dünkelmann, P, and Müller, M (2002). Enantioselective Synthesis of α -Hydroxy Ketones

- via Benzaldehyde Lyase-Catalyzed C-C Bond Formation Reaction. *Adv Synth Catal* **344** (1), 96–103.
- Denger, K, Mayer, J, Buhmann, M, Weinitschke, S, Smits, THM, and Cook, AM (2009). Bifurcated degradative pathway of 3-sulfolactate in *Roseovarius nubinhibens* ISM via sulfoacetaldehyde acetyltransferase and (*S*)-cysteate sulfolyase. *J Bacteriol* **191**, 5648–5656.
- Diefenbach, RJ, Candy, JM, Mattick, JS, and Duggleby, RG (1992). Effects of substitution of aspartate-440 and tryptophan-487 in the thiamin diphosphate binding region of pyruvate decarboxylase from *Zymomonas mobilis*. *FEBS Lett* **296** (1), 95–8.
- Dobritzsch, D, König, S, Schneider, G, and Lu, G (1998). High Resolution Crystal Structure of Pyruvate Decarboxylase from *Zymomonas mobilis*: Implications for Substrate Activation in Pyruvate Decarboxylases. *J Biol Chem* **273** (32), 20196–20204.
- Dodson, G and Wlodawer, A (1998). Catalytic triads and their relatives. *Trends Biochem Sci* **23** (9), 347–352.
- Domínguez de María, P, Pohl, M, Gocke, D, Gröger, H, Trauthwein, H, Stillger, T, Walter, L, and Müller, M (2007). Asymmetric Synthesis of Aliphatic 2-Hydroxy Ketones by Enzymatic Carbonylation of Aldehydes. *European J Org Chem* **2007**.18, 2940–2944.
- Drawz, SM, Bethel, CR, Hujer, KM, Hurless, KN, Distler, AM, Caselli, E, Prati, F, and Bonomo, RA (2009). The Role of a Second-Shell Residue in Modifying Substrate and Inhibitor Interactions in the SHV β -Lactamase: A Study of Ambler Position Asn276. *Biochemistry* **48**.21, 4557–4566.
- Duan, Y, Wu, C, Chowdhury, S, Lee, MC, Xiong, G, Zhang, W, Yang, R, Cieplak, P, Luo, R, Lee, T, Caldwell, J, Wang, J, and Kollman, P (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* **24** (16), 1999–2012.
- Duggleby, RG (2006). Domain relationships in thiamine diphosphate-dependent enzymes. *Acc Chem Res* **39**, 550–557.
- Duggleby, RG and Pang, SS (2000). Acetohydroxyacid synthase. *J Biochem Mol Biol* **33**, 1–36.
- Dünkelmann, P, Kolter-Jung, D, Nitsche, A, Demir, AS, Siegert, P, Linggen, B, Baumann, M, Pohl, M, and Müller, M (2002). Development of a donor-acceptor concept for enzymatic cross-coupling reactions of aldehydes: the first asymmetric cross-benzoin condensation. *J Am Chem Soc* **124**, 12084–12085.

- Dünnwald, T and Müller, M (2000). Stereoselective formation of bis(alpha-hydroxy ketones) via enzymatic carbonylation. *J Org Chem* **65**, 8608–8612.
- Eberhardt, I, Cederberg, H, Li, H, König, S, Jordan, F, and Hohmann, S (1999). Autoregulation of yeast pyruvate decarboxylase gene expression requires the enzyme but not its catalytic activity. *Eur J Biochem* **262** (1), 191–201.
- Edgar, RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461.
- Edwards, MS, Sternberg, MJE, and Thornton, JM (1987). Structural and sequence patterns in the loops of beta-alpha-beta units. *Protein Eng* **1**, 173–181.
- Enders, D and Kallfass, U (2002a). An Efficient Nucleophilic Carbene Catalyst for the Asymmetric Benzoin Condensation. *Angew Chem Int Ed Engl* **41** (10), 1743–1745.
- (2002b). Ein effizienter nucleophiler Carben-Katalysator für die asymmetrische Benzoinkondensation. *Angew Chem Weinheim Bergstr Ger* **114** (10), 1822–1824.
- Engel, S, Vyazmensky, M, Geresh, S, Barak, Z, and Chipman, DM (2003). Acetohydroxyacid synthase: A new enzyme for chiral synthesis of *R*-phenylacetylcarbinol. *Biotechnol Bioeng* **83** (7), 833–840.
- Fang, M, Macova, A, Hanson, KL, Kos, J, and Palmer, DRJ (2011). Using Substrate Analogues To Probe the Kinetic Mechanism and Active Site of *Escherichia coli* MenD. *Biochemistry* **50**, 8712–8721.
- Fang, R, Nixon, PF, and Duggleby, RG (1998). Identification of the catalytic glutamate in the E1 component of human pyruvate dehydrogenase. *FEBS Let* **437** (3), 273–277.
- Fiedler, E, Thorell, S, Sandalova, T, Golbik, R, König, S, and Schneider, G (2002). Snapshot of a key intermediate in enzymatic thiamin catalysis: crystal structure of the alpha-carbanion of (alpha,beta-dihydroxyethyl)-thiamin diphosphate in the active site of transketolase from *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **99**, 591–595.
- Fischer, M, Knoll, M, Sirim, D, Wagner, F, Funke, S, and Pleiss, J (2007). The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. *Bioinformatics* **23** (15), 2015–2017.
- Fischer, M and Pleiss, J (2003). The Lipase Engineering Database: a navigation and analysis tool for protein families. *Nucleic Acids Res* **31** (1), 319–321.
- Fischer, M, Thai, QK, Grieb, M, and Pleiss, J (2006). DWARF - a data warehouse system for analyzing protein families. *BMC Bioinformatics* **7**, 495.

- Fischer, Markus (2004). Die Lipase Engineering Database - Systematische Analyse familienspezifischer Eigenschaften und der Sequenz-Struktur-Funktionsbeziehung von α/β -Hydrolasen. PhD thesis. Institut für Technische Biochemie, Universität Stuttgart, Stuttgart, Germany.
- Fiser, A and Sali, A (2003). Modeller: Generation and Refinement of Homology-Based Protein Structure Models. *Macromolecular Crystallography, Part D*. Ed. by CW Jr. Carter and RM Sweet. **374**. Methods Enzymol. Academic Press, 461–491.
- Fitch, WM (1970). Distinguishing Homologous from Analogous Proteins. *Syst Zool* **19.2**, 99–113.
- Fraccascia, P, Casteels, M, De Schryver, E, and Van Veldhoven, PP (2011). Role of thiamine pyrophosphate in oligomerisation, functioning and import of peroxisomal 2-hydroxyacyl-CoA lyase. *Biochim Biophys Acta* **1814** (10), 1226–1233.
- Fraginello, MC, Hoyos, P, Romano, D, Gandolfi, R, Alcántara, AR, and Molinari, F (2012). Enantioselective reduction and deracemisation using the non-conventional yeast *Pichia glucozyma* in water/organic solvent biphasic systems: preparation of (S)-1,2-diaryl-2-hydroxyethanones (benzoins). *Tetrahedron* **68** (2), 523–528.
- Frank, RAW, Leeper, FJ, and Luisi, BF (2007). Structure, mechanism and catalytic duality of thiamine-dependent enzymes. *Cell Mol Life Sci* **64**, 892–905.
- Frank, RAW, Pratap, JV, Pei, XY, Perham, RN, and Luisi, BF (2005). The molecular origins of specificity in the assembly of a multienzyme complex. *Structure* **13**, 1119–1130.
- Frank, RAW, Price, AJ, Northrop, FD, Perham, RN, and Luisi, BF (2007). Crystal structure of the E1 component of the *Escherichia coli* 2-oxoglutarate dehydrogenase multienzyme complex. *J Mol Biol* **368**, 639–651.
- Frank, RAW, Titman, CM, Pratap, JV, Luisi, BF, and Perham, RN (2004). A Molecular Switch and Proton Wire Synchronize the Active Sites in Thiamine Enzymes. *Science* **306** (5697), 872–876.
- Fries, M, Chauhan, HJ, Domingo, GJ, Jung, HI, and Perham, RN (2003). Site-directed mutagenesis of a loop at the active site of E1 ($\alpha 2\beta 2$) of the pyruvate dehydrogenase complex. *Eur J Biochem* **270** (5), 861–870.
- Fries, M, Jung, HI, and Perham, RN (2003). Reaction Mechanism of the Heterotetrameric ($\alpha 2\beta 2$) E1 Component of 2-Oxo Acid Dehydrogenase Multienzyme Complexes. *Biochemistry* **42** (23), 6996–7002.

- Galleni, M, Lamotte-Brasseur, J, Rossolini, GM, Spencer, J, Dideberg, O, and Frere, JM (2001). Standard numbering scheme for class B beta-lactamases. *Antimicrob Agents Chemother* **45**, 660–663.
- Galman, JL, Steadman, D, Bacon, S, Morris, P, Smith, MEB, Ward, JM, Dalby, PA, and Hailes, HC (2010). α,α' -Dihydroxyketone formation using aromatic and heteroaromatic aldehydes with evolved transketolase enzymes. *Chem Commun* **46** (40), 7608–7610.
- Garau, G, Garcia-Saez, I, Bebrone, C, Anne, C, Mercuri, P, Galleni, M, Frere, JM, and Dideberg, O (2004). Update of the standard numbering scheme for class B beta-lactamases. *Antimicrob Agents Chemother* **48**, 2347–2349.
- Gardossi, L, Poulsen, PB, Ballesteros, A, Hult, K, Svedas, VK, Vasic-Racki, D, Carrea, G, Magnusson, A, Schmid, A, Wohlgemuth, R, and Halling, PJ (2010). Guidelines for reporting of biocatalytic reactions. *Trends Biotechnol* **28.4**, 171–180.
- Gehards, T, Mackfeld, U, Bocola, M, Lieres, E von, Wiechert, W, Pohl, M, and Rother, D (2012). Influence of Organic Solvents on Enzymatic Asymmetric Carbolygations. *Adv Synth Catal* **354**, 2805–2820.
- Gibson, GE, Hirsch, JA, Cirio, RT, Jordan, BD, Fonzetti, P, and Elder, J (2013). Abnormal thiamine-dependent processes in Alzheimer's Disease. Lessons from diabetes. *Mol Cell Neurosci* **55**, 17–25.
- Gocke, D, Graf, T, Brosi, H, Frindi-Wosch, I, Walter, L, Müller, M, and Pohl, M (2009). Comparative characterisation of thiamin diphosphate-dependent decarboxylases. *J Mol Catal B Enzym* **61**, 30–35.
- Gocke, D, Walter, L, Gauchenova, E, Kolter, G, Knoll, M, Berthold, CL, Schneider, G, Pleiss, J, Müller, M, and Pohl, M (2008). Rational protein design of ThDP-dependent enzymes-engineering stereoselectivity. *Chembiochem* **9**, 406–412.
- Gocke, Dörte (2007). New and optimised thiamine diphosphate (ThDP)-dependent enzymes for carbolygation: Creation of a toolbox for chiral 2-hydroxy ketones. PhD thesis. Heinrich-Heine University, Düsseldorf.
- Goetz, G, Iwan, P, Hauer, B, Breuer, M, and Pohl, M (2001). Continuous production of (*R*)-phenylacetylcarbinol in an enzyme-membrane reactor using a potent mutant of pyruvate decarboxylase from *Zymomonas mobilis*. *Biotechnol Bioeng* **74**, 317–325.
- Göttler, Chantal (2013). Identification of structurally equivalent residues in ThDP dependent enzymes. BSc thesis. Institut für Technische Biochemie, Universität Stuttgart.

- Graham, DE, Taylor, SM, Wolf, RZ, and Namboori, SC (2009). Convergent evolution of coenzyme M biosynthesis in the Methanosarcinales: cysteate synthase evolved from an ancestral threonine synthase. *J Mol Biol* **424**, 467–478.
- Graupner, M, Xu, H, and White, RH (2000). Identification of the gene encoding sulfopyruvate decarboxylase, an enzyme involved in biosynthesis of coenzyme M. *J Bacteriol* **182**, 4862–4867.
- Green, JB (1989). Pyruvate decarboxylase is like acetolactate synthase (ILV2) and not like the pyruvate dehydrogenase E1 subunit. *FEBS Lett* **246**, 1–5.
- Gricman, Ł, Vogel, C, and Pleiss, J (2014). Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance. *Proteins* **82**, 491–504.
- Guo, F, Zhang, D, Kahyaoglu, A, Farid, RS, and Jordan, F (1998). Is a hydrophobic amino acid required to maintain the reactive V conformation of thiamin at the active center of thiamin diphosphate-requiring enzymes? Experimental and computational studies of isoleucine 415 of yeast pyruvate decarboxylase. *Biochemistry* **37**, 13379–13391.
- Hailes, HC, Rother, D, Müller, M, Westphal, R, Ward, JM, Pleiss, J, Vogel, C, and Pohl, M (2013). Engineering stereoselectivity of ThDP-dependent enzymes. *FEBS J* **280**, 6374–6394.
- Hallock, MJ, Stone, JE, Roberts, E, Fry, C, and Luthey-Schulten, Z (2014). Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations. *Parallel Comput* **40** (5-6), 86–99.
- Hawkins, CF, Borges, A, and Perham, RN (1989). A common structural motif in thiamin pyrophosphate-binding enzymes. *FEBS Lett* **255**, 77–82.
- Hibbert, EG, Senussi, T, Smith, MEB, Costelloe, SJ, Ward, JM, Hailes, HC, and Dalby, PA (2008). Directed evolution of transketolase substrate specificity towards an aliphatic aldehyde. *J Biotechnol* **134** (3-4), 240–245.
- Hildebrandt, G and Klavehn, W. (1930). Verfahren zur Herstellung von 1-L-Phenyl-2-methylaminopropan-1-ol. Knoll A.-G. Chemische Fabriken in Ludwigshafen.
- Hill, CM and Duggleby, RG (1998). Mutagenesis of *Escherichia coli* acetohydroxyacid synthase isoenzyme II and characterization of three herbicide-insensitive forms. *Biochem J* **335** (3), 653–661.
- HMMER, <http://hmmer.janelia.org/> (2013).
- Honegger, A and Plückthun, A (2001). Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool. *J Mol Biol* **309** (3), 657–670.

- Hotelier, T, Renault, L, Cousin, X, Negre, V, Marchot, P, and Chatonnet, A (2004). ESTHER, the database of the α/β -hydrolase fold superfamily of proteins. *Nucleic Acids Res* **32**, D145–D147.
- Hoyos, P, Sinisterra, JV, Molinari, F, Alcántara, AR, and María, PD de (2010). Biocatalytic Strategies for the Asymmetric Synthesis of α -Hydroxy Ketones. *Acc Chem Res* **43** (2), 288–299.
- Huang, CY, Chang, AK, Nixon, PF, and Duggleby, RG (2001). Site-directed mutagenesis of the ionizable groups in the active site of *Zymomonas mobilis* pyruvate decarboxylase: effect on activity and pH dependence. *Eur J Biochem* **268**, 3558–3565.
- Iding, H, Dünnwald, T, Greiner, L, Liese, A, Müller, M, Siegert, P, Grötzinger, J, Demir, AS, and Pohl, M (2000). Benzoylformate Decarboxylase from *Pseudomonas putida* as Stable Catalyst for the Synthesis of Chiral 2-Hydroxy Ketones. *Chemistry* **6** (8), 1483–1495.
- Iding, H, Siegert, P, and Pohl, M (1998). Application of α -keto acid decarboxylases in biotransformations. *Biochim Biophys Acta* **1385** (2), 307–322.
- Iwan, P, Goetz, G, Schmitz, S, Hauer, B, Breuer, M, and Pohl, M (2001). Studies on the continuous production of (*R*)-(-)-phenylacetylcarbinol in an enzyme-membrane reactor. *J Mol Catal B Enzym* **11**, 387–396.
- Janzen, E, Müller, M, Kolter-Jung, D, Kneen, MM, McLeish, MJ, and Pohl, M (2006). Characterization of benzaldehyde lyase from *Pseudomonas fluorescens* - a versatile enzyme for asymmetric C-C-bond formation. *Bioorg Chem* **34**, 345–361.
- Jiang, M, Cao, Y, Guo, Z-F, Chen, M, Chen, X, and Guo, Z (2007). Menaquinone Biosynthesis in *Escherichia coli*: Identification of 2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate as a Novel Intermediate and Re-Evaluation of MenD Activity. *Biochemistry* **46**, 10979–10989.
- Johnen, S and Sprenger, GA (2009). Characterization of recombinant thiamine diphosphate-dependent phosphonopyruvate decarboxylase from *Streptomyces viridochromogenes* Tü494. *J Mol Catal B* **61**, 39–46.
- Johnson, LS, Eddy, SR, and Portugaly, E (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431.
- Joosten, Hendrik Johannes (2007). 3DM: From Data to Medicine. PhD thesis. Wageningen University, Wageningen, Netherlands.
- Jordan, F (2003). Current mechanistic understanding of thiamin diphosphate-dependent enzymatic reactions. *Nat Prod Rep* **20**, 184–201.

- Jordan, F, Li, H, and Brown, A (1999). Remarkable stabilization of zwitterionic intermediates may account for a billion-fold rate acceleration by thiamin diphosphate-dependent decarboxylases. *Biochemistry* **38**, 6369–6373.
- Jordan, F, Nemeria, N, Guo, F, Baburina, I, Gao, Y, Kahyaoglu, A, Li, H, Wang, J, Yi, J, Guest, JR, and Furey, W (1998). Regulation of thiamin diphosphate-dependent 2-oxo acid decarboxylases by substrate and thiamin diphosphate.Mg(II) - evidence for tertiary and quaternary interactions. *Biochim Biophys Acta* **1385** (2), 287–306.
- Jordan, F, Zhang, Z, and Sergienko, E (2002). Spectroscopic Evidence for Participation of the 1',4'-Imino Tautomer of Thiamin Diphosphate in Catalysis by Yeast Pyruvate Decarboxylase. *Bioorg Chem* **30** (3), 188–198.
- Jorgensen, WL, Maxwell, DS, and Tirado-Rives, J (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J Am Chem Soc* **118** (45), 11225–11236.
- Joseph, E, Wei, W, Tittmann, K, and Jordan, F (2006). Function of a conserved loop of the beta-domain, not involved in thiamin diphosphate binding, in catalysis and substrate activation in yeast pyruvate decarboxylase. *Biochemistry* **45**, 13517–13527.
- Kabat, EA, Wu, TT, and Perry, H (1991). Sequences of Proteins of Immunological Interest. Fifth Edition. Bethesda, MD: NIH Publication.
- Kakrana, A, Hammond, R, Patel, P, Nakano, M, and Meyers, BC (2014). *sPARTA*: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Res*. doi: [10.1093/nar/gku693](https://doi.org/10.1093/nar/gku693).
- Kaminuma, E, Kosuge, T, Kodama, Y, Aono, H, Mashima, J, Gojobori, T, Sugawara, H, Ogasawara, O, Takagi, T, Okubo, K, and Nakamura, Y (2011). DDBJ progress report. *Nucleic Acids Res* **39**, D22–27.
- Kanehisa, M, Goto, S, Hattori, M, Aoki-Kinoshita, KF, Itoh, M, Kawashima, S, Katayama, T, Araki, M, and Hirakawa, M (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34**, D354–D357.
- Kaplun, A, Binshtein, E, Vyazmensky, M, Steinmetz, A, Barak, Z, Chipman, DM, Tittmann, K, and Shaanan, B (2008). Glyoxylate carboligase lacks the canonical active site glutamate of thiamine-dependent enzymes. *Nat Chem Biol* **4**, 113–118.

- Kara, S, Long, WS, Berheide, M, Peper, S, Niemeyer, B, and Liese, A (2011). Influence of reaction conditions on the enantioselectivity of biocatalyzed C-C bond formations under high pressure conditions. *J Biotechnol* **152** (3), 87–92.
- Karulin, B and Kozhevnikov, M (2011). Ketcher: web-based chemical structure editor. *J Cheminform* **3**.Suppl 1, P3.
- Kaziyo, Y, Tanaka, R, Mano, Y, and Shimazono, N (1961). On the Mechanism of Transpyrophosphorylation in the Biosynthesis of Thiamine Diphosphate. *J Biochem* **49** (6), 472–476.
- Kern, D, Kern, G, Neef, H, Tittmann, K, Killenberg-Jabs, M, Wikner, C, Schneider, G, and Hübner, G (1997). How Thiamine Diphosphate Is Activated in Enzymes. *Science* **275** (5296), 67–70.
- Killenberg-Jabs, M, König, S, Eberhardt, I, Hohmann, S, and Hübner, G (1997). Role of Glu51 for Cofactor Binding and Catalytic Activity in Pyruvate Decarboxylase from Yeast Studied by Site-Directed Mutagenesis. *Biochemistry* **36**, 1900–1905.
- Kluger, R and Tittmann, K (2008). Thiamin diphosphate catalysis: enzymic and nonenzymic covalent intermediates. *Chem Rev* **108**, 1797–1833.
- Kneen, MM, Pogozheva, ID, Kenyon, GL, and McLeish, MJ (2005). Exploring the active site of benzaldehyde lyase by modeling and mutagenesis. *Biochim Biophys Acta* **1753**, 261–271.
- Kneen, MM, Stan, R, Yep, A, Tyler, RP, Saehuan, C, and McLeish, MJ (2011). Characterization of a thiamin diphosphate-dependent phenylpyruvate decarboxylase from *Saccharomyces cerevisiae*. *FEBS J* **278** (11), 1842–1853.
- Knoll, M, Hamm, T, Wagner, F, Martinez, V, and Pleiss, J (2009). The PHA Depolymerase Engineering Database: A systematic analysis tool for the diverse family of polyhydroxyalkanoate (PHA) depolymerases. *BMC Bioinformatics* **10** (1), 89.
- Knoll, M, Müller, M, Pleiss, J, and Pohl, M (2006). Factors mediating activity, selectivity, and substrate specificity for the thiamin diphosphate-dependent enzymes benzaldehyde lyase and benzoylformate decarboxylase. *Chembiochem* **7**, 1928–1934.
- Knoll, M and Pleiss, J (2008). 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.
- Kochetov, GA and Solovjeva, ON (2014). Structure and functioning mechanism of transketolase. *Biochim Biophys Acta* **1844**.9, 1608–1618.

- Koga, J, Adachi, T, and Hidaka, H (1991). Molecular cloning of the gene for indolepyruvate decarboxylase from *Enterobacter cloacae*. *Mol Gen Genet* **226**, 10–16.
- Kokova, M, Zavrel, M, Tittmann, K, Spiess, AC, and Pohl, M (2009). Investigating the carboli-gase activity of thiamine diphosphate-dependent enzymes using kinetic modeling and NMR spectroscopy. *Biotrans conference*.
- Kosloff, M and Kolodny, R (2008). Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins* **71** (2), 891–902.
- Kowarsch, A, Fuchs, A, Frishman, D, and Pagel, P (2010). Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions. *PLoS Comput Biol* **6.9**, e1000923.
- Krawczyk, E, Koprowski, M, Skowrońska, A, and Łuczak, J (2004). α -Hydroxy ketones in high enantiomeric purity from asymmetric oxidation of enol phosphates with (salen) manganese(III) complex. *Tetrahedron Asymmetry* **15** (17), 2599–2602.
- Krieger, E, Koraimann, G, and Vriend, G (2002). Increasing the precision of comparative models with YASARA NOVA -a self-parameterizing force field. *Proteins* **47** (3), 393–402.
- Kuipers, RK, Joosten, HJ, Verwiel, E, Paans, S, Akerboom, J, Oost, J van der, Lefterink, NG, Berkel, WJ van, Vriend, G, and Schaap, PJ (2009). Correlated mutation analyses on super-family alignments reveal functionally important residues. *Proteins* **76.3**, 608–616.
- Kutter, S, Weiss, MS, Wille, G, Golbik, R, Spinka, M, and König, S (2009). Covalently bound substrate at the regulatory site of yeast pyruvate decarboxylases triggers allosteric enzyme activation. *J Biol Chem* **284**, 12136–12144.
- Larkin, MA, Blackshields, G, Brown, NP, Chenna, R, McGettigan, PA, McWilliam, H, Valentin, F, Wallace, IM, Wilm, A, Lopez, R, Thompson, JD, Gibson, TJ, and Higgins, DG (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23** (21), 2947–2948.
- Al-Lazikani, B, Lesk, AM, and Chothia, C (1997). Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* **273**, 927–948.
- Le, DT, Yoon, MY, Kim, YT, and Choi, JD (2004). Homology modeling of the structure of tobacco acetohydroxy acid synthase and examination of the active site by site-directed mutagenesis. *Biochem Biophys Res Commun* **317** (3), 930–938.
- (2005). Roles of Three Well-Conserved Arginine Residues in Mediating the Catalytic Activity of Tobacco Acetohydroxy Acid Synthase. *J Biochem* **138** (1), 35–40.

- Lee, MY, Lee, SC, Cho, JH, Ryu, SE, Koo, BS, and Yoon, MY (2013). Role of a highly conserved and catalytically important glutamate-49 in the *Enterococcus faecalis* acetolactate synthase. *Bull Korean Chem Soc* **34**, 669–672.
- Lee, SC, Jung, IP, Baig, IA, Chien, PN, La, IJ, and Yoon, MY (2015). Mutational analysis of critical residues of FAD-independent catabolic acetolactate synthase from *Enterococcus faecalis* {V583}. *Int J Biol Macromol* **72**, 104–109.
- Lehmann, M, Loch, C, Middendorf, A, Studer, D, Lassen, SF, Pasamontes, L, Loon, A PGM van, and Wyss, M (2002). The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng* **15.5**, 403–411.
- Lerner, RA, Benkovic, SJ, and Schultz, PG (1991). At the crossroads of chemistry and immunology: catalytic antibodies. *Science* **252** (5006), 659–667.
- Li, H, Furey, W, and Jordan, F (1999). Role of glutamate 91 in information transfer during substrate activation of yeast pyruvate decarboxylase. *Biochemistry* **38**, 9992–10003.
- Li, H and Jordan, F (1999). Effects of substitution of tryptophan 412 in the substrate activation pathway of yeast pyruvate decarboxylase. *Biochemistry* **38**, 10004–10012.
- Lie, MA, Celik, L, Jørgensen, KA, and Schiøtt, B (2005). Cofactor Activation and Substrate Binding in Pyruvate Decarboxylase. Insights into the Reaction Mechanism from Molecular Dynamics Simulations. *Biochemistry* **44** (45), 14792–14806.
- Linderstrøm-Lang, KU (1952). Lane Medial Lectures: Proteins and Enzymes. *Proteins and Enzymes*. **6**. Lane Medial Lectures. Stanford University Press, Stanford, California.
- Lindqvist, Y and Schneider, G (1993). Thiamin Diphosphate Dependent Enzymes - Transketolase, Pyruvate Oxidase and Pyruvate Decarboxylase. *Curr Opin Biotechnol* **3**, 896–901.
- Lingen, B, Grötzinger, J, Kolter, D, Kula, MR, and Pohl, M (2002). Improving the carbonylase activity of benzoylformate decarboxylase from *Pseudomonas putida* by a combination of directed evolution and site-directed mutagenesis. *Protein Eng* **15** (7), 585–593.
- Lingen, B, Kolter-Jung, D, Dunkelmann, P, Feldmann, R, Grötzinger, J, Pohl, M, and Müller, M (2003). Alteration of the substrate specificity of benzoylformate decarboxylase from *Pseudomonas putida* by directed evolution. *Chembiochem* **4**, 721–726.
- Lipman, DJ and Pearson, WR (1985). Rapid and sensitive protein similarity searches. *Science* **227** (4693), 1435–1441.
- Liu, M, Sergienko, EA, Guo, F, Wang, J, Tittmann, K, Hübner, G, Furey, W, and Jordan, F (2001). Catalytic acid-base groups in yeast pyruvate decarboxylase. 1. Site-directed mutagenesis

- and steady-state kinetic studies on the enzyme with the D28A, H114F, H115F, and E477Q substitutions. *Biochemistry* **40**, 7355–7368.
- Lobell, M and Crout, DHG (1996). Pyruvate decarboxylase: A molecular modeling study of pyruvate decarboxylation and acyloin formation. *J Am Chem Soc* **118**, 1867–1873.
- Lombard, V, Golaconda Ramulu, H, Drula, E, Coutinho, PM, and Henrissat, B (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**.D1, D490–D495.
- Loschonsky, Sabrina (2014). Untersuchungen zur C-C-Bindungsknüpfungs- und C-C-Bindungsspaltungs-Aktivität des Thiamindiphosphat-abhängigen Enzyms Cyclohexan-1,2-dion Hydrolase. PhD thesis. Albert-Ludwigs-Universität Freiburg, Freiburg.
- Lu, G, Dobritsch, D, Baumann, S, Schneider, G, and König, S (2000). The structural basis of substrate activation in yeast pyruvate decarboxylase. A crystallographic and kinetic study. *Eur J Biochem* **267**, 861–868.
- Lu'o'ng, Kv and Nguyen, LT (2011). Role of Thiamine in Alzheimer's Disease. *Am J Alzheimers Dis Other Demen* **26** (8), 588–598.
- Majorek, KA, Kuhn, ML, Chruszcz, M, Anderson, WF, and Minor, W (2014). Double trouble- Buffer selection and His-tag presence may be responsible for nonreproducibility of biomedical experiments. *Protein Sci* **23**.10, 1359–1368.
- Mazin, P, Gelfand, M, Mironov, A, Rakhmaninova, A, Rubinov, A, Russell, R, and Kalinina, O (2010). An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. *Algorithms Mol Biol* **5**.1, 29.
- McCourt, JA, Pang, SS, King-Scott, J, Guddat, LW, and Duggleby, RG (2006). Herbicide-binding sites revealed in the structure of plant acetohydroxyacid synthase. *Proc Natl Acad Sci USA* **103**, 569–573.
- McDonald, A, Boyce, S, Moss, G, Dixon, H, and Tipton, K (2007). ExplorEnz: a MySQL database of the IUBMB enzyme nomenclature. *BMC Biochem* **8**.1, 14.
- Meunier, B, Visser, SP de, and Shaik, S (2004). Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes. *Chem Rev* **104** (9), 3947–3980.
- Meyer, D, Neumann, P, Ficner, R, and Tittmann, K (2013). Observation of a stable carbene at the active site of a thiamin enzyme. *Nat Chem Biol* **9**, 488–490.
- Meyer, D, Neumann, P, Koers, E, Sjuts, H, Lüdtkke, S, Sheldrick, GM, Ficner, R, and Tittmann, K (2012). Unexpected tautomeric equilibria of the carbanion-enamine intermediate in pyruvate

- oxidase highlight unrecognized chemical versatility of thiamin. *Proc Natl Acad Sci USA* **109**, 10867–10872.
- Meyer, D, Neumann, P, Parthier, C, Friedemann, R, Nemeria, N, Jordan, F, and Tittmann, K (2010). Double duty for a conserved glutamate in pyruvate decarboxylase: evidence of the participation in stereoelectronically controlled decarboxylation and in protonation of the nascent carbanion/enamine intermediate. *Biochemistry* **49**, 8197–8212.
- Meyer, D, Walter, L, Kolter, G, Pohl, M, Müller, M, and Tittmann, K (2011). Conversion of pyruvate decarboxylase into an enantioselective carboligase with biosynthetic potential. *J Am Chem Soc* **133**, 3609–3616.
- Meyer, Danilo (2009). Kinetische und strukturelle Untersuchung der Katalysemechanismen ausgewählter Kofaktor-abhängiger Enzyme - Implikationen für die Decarboxylierung von α -Ketosäuren durch Thiamindiphosphat-abhängige Enzyme. PhD thesis. Martin-Luther-Universität, Halle-Wittenberg.
- Moody, Glyn (2004). Digital code of life: How bioinformatics is revolutionizing science, medicine, and business. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 11–33.
- Mosbacher, TG, Müller, M, and Schulz, GE (2005). Structure and mechanism of the ThDP-dependent benzaldehyde lyase from *Pseudomonas fluorescens*. *FEBS J* **272** (23), 6067–6076.
- Müller, M, Gocke, D, and Pohl, M (2009). Thiamin diphosphate in biological chemistry: exploitation of diverse thiamin diphosphate-dependent enzymes for asymmetric chemoenzymatic synthesis. *FEBS J* **276**, 2894–2904.
- Müller, M, Kurutsch, A, Richter, M, Brecht, V, and Sprenger, GA (2009). MenD as a versatile catalyst for asymmetric synthesis. *J Mol Catal B Enzym* **61**, 56–66.
- Müller, M, Sprenger, GA, and Pohl, M (2013). C-C bond formation using ThDP-dependent lyases. *Curr Opin Chem Biol* **17**, 261–270.
- Muller, YA, Lindqvist, Y, Furey, W, Schulz, GE, Jordan, F, and Schneider, G (1993). A thiamin diphosphate binding fold revealed by comparison of the crystal structures of transketolase, pyruvate oxidase and pyruvate decarboxylase. *Structure* **1**, 95–103.
- Nelson, D (2009). The Cytochrome P450 Homepage. *Hum Genomics* **4.1**, 59–65.
- Nemeria, N, Binshtein, E, Patel, H, Balakrishnan, A, Vered, I, Shaanan, B, Barak, Z, Chipman, D, and Jordan, F (2012). Glyoxylate Carboligase: A Unique Thiamin Diphosphate-Dependent Enzyme That Can Cycle between the 4'-Aminopyrimidinium and 1',4'-Iminopyrimidine

- Tautomeric Forms in the Absence of the Conserved Glutamate. *Biochemistry* **51** (40), 7940–7952.
- Nemeria, N, Tittmann, K, Joseph, E, Zhou, L, Vazquez-Coll, MB, Arjunan, P, Hübner, G, Furey, W, and Jordan, F (2005). Glutamate 636 of the *Escherichia coli* Pyruvate Dehydrogenase-E1 Participates in Active Center Communication and Behaves as an Engineered Acetolactate Synthase with Unusual Stereoselectivity. *J Biol Chem* **280** (22), 21473–21482.
- Nemeria, N, Volkov, A, Brown, A, Yi, J, Zipper, L, Guest, JR, and Jordan, F (1998). Systematic Study of the Six Cysteines of the E1 Subunit of the Pyruvate Dehydrogenase Multienzyme Complex from *Escherichia coli*: None Is Essential for Activity. *Biochemistry* **37** (3), 911–922.
- Nemeria, N, Yan, Y, Zhang, Z, Brown, AM, Arjunan, P, Furey, W, Guest, JR, and Jordan, F (2001). Inhibition of the *Escherichia coli* Pyruvate Dehydrogenase Complex E1 Subunit and Its Tyrosine 177 Variants by Thiamin 2-Thiazolone and Thiamin 2-Thiothiazolone Diphosphates: Evidence for reversible tight-binding inhibition. *J Biol Chem* **276** (49), 45969–45978.
- Nemeria, NS, Arjunan, P, Chandrasekhar, K, Mossad, M, Tittmann, K, Furey, W, and Jordan, F (2010). Communication between Thiamin Cofactors in the *Escherichia coli* Pyruvate Dehydrogenase Complex E1 Component Active Centers: evidence for a 'direct pathway' between the 4' -aminopyrimidine N1' atoms. *J Biol Chem* **285** (15), 11197–11209.
- Neuberg, C and Hirsch, J (1921). Über ein Kohlenstoffketten knüpfendes Ferment (Carbologase). *Biochem Z* **115**, 282–310.
- Notredame, C, Higgins, DG, and Heringa, J (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205–217.
- O'Reilly, M, Watson, KA, and Johnson, LN (1999). The Crystal Structure of the *Escherichia coli* Maltodextrin Phosphorylase-Acarbose Complex. *Biochemistry* **38** (17), 5337–5345.
- Orengo, CA, Jones, DT, and Thornton, JM (1994). Protein superfamilies and domain superfolds. *Nature* **372**, 631–634.
- Pang, SS, Duggleby, RG, Schowen, RL, and Guddat, LW (2004). The crystal structures of *Klebsiella pneumoniae* acetolactate synthase with enzyme-bound cofactor and with an unusual intermediate. *J Biol Chem* **279**, 2242–2253.
- Payongsri, P, Steadman, D, Strafford, J, MacMurray, A, Hailes, HC, and Dalby, PA (2012). Rational substrate and enzyme engineering of transketolase for aromatics. *Org Biomol Chem* **10**, 9021–9029.

- Pearson, WR and Lipman, DJ (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85** (8), 2444–2448.
- Pei, XY, Titman, CM, Frank, RAW, Leeper, FJ, and Luisi, BF (2008). Snapshots of Catalysis in the E1 Subunit of the Pyruvate Dehydrogenase Multienzyme Complex. *Structure* **16** (12), 1860–1872.
- Pleiss, J, Fischer, M, Peiker, M, Thiele, C, and Schmid, RD (2000). Lipase engineering database: Understanding and exploiting sequence-structure-function relationships. *J Mol Catal B Enzym* **10** (5), 491–508.
- Pletcher, J, Wood, M, Blank, G, Shin, W, and Sax, M (1977). Thiamine pyrophosphate tetrahydrate: a structure with the pyrophosphate ester in an extended conformation. *Acta Crystallogr Sect B Struct Crystallogr Cryst Chem* **33**, 3349–3359.
- Pohl, M (1997). Protein design on pyruvate decarboxylase (PDC) by site-directed mutagenesis. Application to mechanistical investigations, and tailoring PDC for the use in organic synthesis. *Adv Biochem Eng Biotechnol* **58**, 15–43.
- Pohl, M, Dresen, C, Beigi, M, and Müller, M (2012). Enzymatic Acyloin and Benzoin Condensations. Ed. by K Drauz, H Gröger, and O May. Weinheim: Wiley-VCH, 919–945.
- Pohl, M, Gocke, D, and Müller, Michael (2010). Thiamine-Based Enzymes for Biotransformations. Ed. by PT Anastat. Wiley-VCH Verlag GmbH & Co. KGaA, 75–114.
- Pohl, M, Lingen, B, and Müller, M (2002). Thiamin-diphosphate-dependent enzymes: new aspects of asymmetric C-C bond formation. *Chemistry* **8**, 5288–5295.
- Pohl, M, Siegert, P, Mesch, K, Bruhn, H, and Grötzinger, J (1998). Active site mutants of pyruvate decarboxylase from *Zymomonas mobilis* - A site-directed mutagenesis study of L112, I472, I476, E473 and N482. *Eur J Biochem* **257**, 538–546.
- Pohl, M, Sprenger, GA, and Müller, M (2004). A new perspective on thiamine catalysis. *Curr Opin Biotechnol* **15**, 335–342.
- Polovnikova, ES, McLeish, MJ, Sergienko, EA, Burgner, JT, Anderson, NL, Bera, AK, Jordan, F, Kenyon, GL, and Hasson, MS (2003). Structural and Kinetic Analysis of Catalysis by a Thiamin Diphosphate-Dependent Enzyme, Benzoylformate Decarboxylase. *Biochemistry* **42** (7), 1820–1830.
- Pronk, S, Páll, S, Schulz, R, Larsson, P, Bjelkmar, P, Apostolov, R, Shirts, MR, Smith, JC, Kasson, PM, Spoel, D van der, Hess, B, and Lindahl, E (2013). GROMACS 4.5: a high-

- throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29** (7), 845–854.
- Querol, J, Rodríguez-Concepción, M, Boronat, A, and Imperial, S (2001). Essential Role of Residue H49 for Activity of *Escherichia coli* 1-Deoxy-D-xylulose 5-Phosphate Synthase, the Enzyme Catalyzing the First Step of the 2-C-Methyl-D-erythritol 4-Phosphate Pathway for Isoprenoid Synthesis. *Biochem Biophys Res Commun* **289** (1), 155–160.
- Racolta, S, Juhl, PB, Sirim, D, and Pleiss, J (2012). The triterpene cyclase protein family: A systematic analysis. *Proteins* **80** (8), 2009–2019.
- Ranoux, A, Karmee, SK, Jin, J, Bhaduri, A, Caiazzo, A, Arends, IWCE, and Hanefeld, U (2012). Enhancement of the Substrate Scope of Transketolase. *Chembiochem* **13** (13), 1921–1931.
- Reinert, DJ, Balliano, G, and Schulz, GE (2004). Conversion of Squalene to the Pentacarboxylic Hopene. *Chem Biol* **11** (1), 121–126.
- Rice, P, Longden, I, and Bleasby, A (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet* **16** (6), 23.
- Rosche, B, Breuer, M, Hauer, B, and Rogers, PL (2005). Role of pyruvate in enhancing pyruvate decarboxylase stability towards benzaldehyde. *J Biotechnol* **115**, 91–99.
- Rost, B (1999). Twilight zone of protein sequence alignments. *Protein Eng* **12** (2), 85–94.
- Rother, D, Kolter, G, Gerhards, T, Berthold, CL, Gauchenova, E, Knoll, M, Pleiss, J, Müller, M, Schneider, G, and Pohl, M (2011). S-Selective mixed carbonylation by structure-based design of the pyruvate decarboxylase from *Acetobacter pasteurianus*. *ChemCatChem* **3**, 1587–1596.
- Russell, RB (2002). Classification of protein folds. *Mol Biotechnol* **20** (1), 17–28.
- Russell, RB and Barton, GJ (1992). Multiple Protein-Sequence Alignment from Tertiary Structure Comparison - Assignment of Global and Residue Confidence Levels. *Proteins* **14**, 309–323.
- Sali, A and Blundell, TL (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J Mol Biol* **234** (3), 779–815.
- Sambrook, J and Russel, DW (2001). *Molecular Cloning: A Laboratory Manual*. First Edition. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, New York, 1116–1118.
- Schellenberger, A (1998). Sixty years of thiamin diphosphate biochemistry. *Biochim Biophys Acta* **1385.2**, 177–186.
- Scheller, PN, Fademrecht, S, Hofelzer, S, Pleiss, J, Leipold, F, Turner, NJ, Nestl, BM, and Hauer, B (2014). Enzyme Toolbox: Novel Enantiocomplementary Imine Reductases. *ChemBioChem* **15.15**, 2201–2204.

- Schenk, G, Duggleby, RG, and Nixon, PF (1998). Properties and functions of the thiamin diphosphate dependent enzyme transketolase. *Int J Biochem Cell Biol* **30**, 1297–1318.
- Schenk, G, Leeper, FJ, England, R, Nixon, PF, and Duggleby, RG (1997). The role of His113 and His114 in pyruvate decarboxylase from *Zymomonas mobilis*. *Eur J Biochem* **248**, 63–71.
- Schmitz, Carlo (2012). Untersuchung der Stereoselektivität und des Substratspektrums von AHAS I & II aus *Escherichia coli*. BSc thesis. Fachhochschule Aachen.
- Schomburg, I, Chang, A, Placzek, S, Söhngen, C, Rother, M, Lang, M, Munaretto, C, Ulas, S, Stelzer, M, Grote, A, Scheer, M, and Schomburg, D (2013). BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res* **41**.D1, D764–D772.
- Schrag, JD, Li, Y, Wu, S, and Cygler, M (1991). Ser-His-Glu triad forms the catalytic site of the lipase from *Geotrichum candidum*. *Nature* **351**, 761–764.
- Schütz, A, Golbik, R, König, S, Hübner, G, and Tittmann, K (2005). Intermediates and Transition States in Thiamin Diphosphate-Dependent Decarboxylases. A Kinetic and NMR Study on Wild-Type Indolepyruvate Decarboxylase and Variants Using Indolepyruvate, Benzoylformate, and Pyruvate as Substrates. *Biochemistry* **44** (16), 6164–6179.
- Schütz, A, Sandalova, T, Ricagno, S, Hübner, G, König, S, and Schneider, G (2003). Crystal structure of thiamindiphosphate-dependent indolepyruvate decarboxylase from *Enterobacter cloacae*, an enzyme involved in the biosynthesis of the plant hormone indole-3-acetic acid. *Eur J Biochem* **270**, 2312–2321.
- Selivanov, VA, Kovina, MV, Kochevova, NV, Meshalkina, LE, and Kochetov, GA (2004). Kinetic study of the H103A mutant yeast transketolase. *FEBS Lett* **567** (2-3), 270–274.
- Sergienko, EA and Jordan, F (2001a). Catalytic acid-base groups in yeast pyruvate decarboxylase. 2. Insights into the specific roles of D28 and E477 from the rates and stereospecificity of formation of carboligase side products. *Biochemistry* **40**, 7369–7381.
- (2001b). Catalytic acid-base groups in yeast pyruvate decarboxylase. 3. A steady-state kinetic model consistent with the behavior of both wild-type and variant enzymes at all relevant pH values. *Biochemistry* **40**, 7382–7403.
- Sergienko, EA, Wang, J, Polovnikova, L, Hasson, MS, McLeish, MJ, Kenyon, GL, and Jordan, F (2000). Spectroscopic Detection of Transient Thiamin Diphosphate-Bound Intermediates on Benzoylformate Decarboxylase. *Biochemistry* **39** (45), 13862–13869.

- Shaanan, B and Chipman, DM (2009). Reaction mechanisms of thiamin diphosphate enzymes: new insights into the role of a conserved glutamate residue. *FEBS J* **276**, 2447–2453.
- Shaw, DE, Deneroff, MM, Dror, RO, Kuskin, JS, Larson, RH, Salmon, JK, Young, C, Batson, B, Bowers, KJ, Chao, JC, Eastwood, MP, Gagliardo, J, Grossman, JP, Ho, CR, Ierardi, DJ, Kolossváry, I, Klepeis, JL, Layman, T, McLeavey, C, Moraes, MA, Mueller, R, Priest, EC, Shan, Y, Spengler, J, Theobald, M, Towles, B, and Wang, SC (2008). Anton, a Special-purpose Machine for Molecular Dynamics Simulation. *Commun ACM* **51** (7), 91–97.
- Shen, MY and Sali, A (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15** (11), 2507–2524.
- Shetty, RP, Bakker, PIW de, DePristo, MA, and Blundell, TL (2003). Advantages of fine-grained side chain conformer libraries. *Protein Eng* **16** (12), 963–969.
- Shin, W, Pletcher, J, Blank, G, and Sax, M (1977). Ring stacking interactions between thiamin and planar molecules as seen in the crystal structure of a thiamin picrolonate dihydrate complex. *J Am Chem Soc* **99** (10), 3491–3499.
- Siegert, P, McLeish, MJ, Baumann, M, Iding, H, Kneen, MM, Kenyon, GL, and Pohl, M (2005). Exchanging the substrate specificities of pyruvate decarboxylase from *Zymomonas mobilis* and benzoylformate decarboxylase from *Pseudomonas putida*. *Protein Eng Des Sel* **18**, 345–357.
- Siegert, Petra (2000). Vergleichende Charakterisierung der Decarboxylase- und Carboligasereaktion der Benzoylformiatdecarboxylase aus *Pseudomonas putida* und der Pyruvatdecarboxylase aus *Zymomonas mobilis* mittels gerichteter Mutagenese. PhD thesis. Heinrich-Heine University, Düsseldorf.
- Sievers, F, Wilm, A, Dineen, D, Gibson, TJ, Karplus, K, Li, WZ, Lopez, R, McWilliam, H, Remmert, M, and Soding, J (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539.
- Sirim, D, Wagner, F, Lisitsa, A, and Pleiss, J (2009). The Cytochrome P450 Engineering Database: integration of biochemical properties. *BMC Biochemistry* **10** (1), 27.
- Sirim, D, Wagner, F, Wang, L, Schmid, RD, and Pleiss, J (2011). The Laccase Engineering Database: a classification and analysis system for laccases and related multicopper oxidases. *Database (Oxford)* **bar006**.
- Sirim, D, Widmann, M, Wagner, F, and Pleiss, J (2010). Prediction and analysis of the modular structure of cytochrome P450 monooxygenases. *BMC Struct Biol* **10.1**, 34.

- Smith, MEB, Hibbert, EG, Jones, AB, Dalby, PA, and Hailes, HC (2008). Enhancing and Reversing the Stereoselectivity of *Escherichia coli* Transketolase via Single-Point Mutations. *Adv Synth Catal* **350** (16), 2631–2638.
- Smoot, ME, Ono, K, Ruscheinski, J, Wang, PL, and Ideker, T (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432.
- Soh, Y, Song, BJ, Jeng, J, and Kallarakal, AT (1998). Critical role of Arg433 in rat transketolase activity as probed by site-directed mutagenesis. *Biochem J* **333** (2), 367–372.
- Spoel, D aan der, Lindahl, E, Hess, B, Groenhof, G, Mark, AE, and Berendsen, HJC (2005). GROMACS: Fast, flexible, and free. *J Comput Chem* **26** (16), 1701–1718.
- Steinbach, A, Fraas, S, Harder, J, Warkentin, E, Kroneck, PMH, and Ermler, U (2012). Crystal structure of a ring-cleaving cyclohexane-1,2-dione hydrolase, a novel member of the thiamine diphosphate enzyme family. *FEBS J* **279**, 1209–1219.
- Steinmetz, A, Vyazmensky, M, Meyer, D, Barak, Z, Golbik, R, Chipman, DM, and Tittmann, K (2010). Valine 375 and Phenylalanine 109 Confer Affinity and Specificity for Pyruvate as Donor Substrate in Acetohydroxy Acid Synthase Isozyme II from *Escherichia coli*. *Biochemistry* **49** (25), 5188–5199.
- Stierand, K and Rarey, M (2007). From Modeling to Medicinal Chemistry: Automatic Generation of Two-Dimensional Complex Diagrams. *ChemMedChem* **2** (6), 853–860.
- Strasser, BJ (2010). Collecting, Comparing, and Computing Sequences: The Making of Margaret O. Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965. *J Hist Biol* **43** (4), 623–660.
- Suplatov, D, Panin, N, Kirilin, E, Shcherbakova, T, Kudryavtsev, P, and Svedas, V (2014). Computational Design of a pH Stable Enzyme: Understanding Molecular Mechanism of Penicillin Acylase's Adaptation to Alkaline Conditions. *PLoS ONE* **9**.6, e100643.
- Suplatov, DA, Besenmatter, W, Svedas, VK, and Svendsen, A (2012). Bioinformatic analysis of alpha/beta-hydrolase fold enzymes reveals subfamily-specific positions responsible for discrimination of amidase and lipase activities. *Protein Eng Des Sel* **25**.11, 689–697.
- Takenaka, A, Juan, ECM, Hoque, MM, Hossain, MT, Yamamoto, T, Imamura, S, Suzuki, K, and Sekiguchi, T (2007). The structures of pyruvate oxidase from *Aerococcus viridans* with cofactors and with a reaction intermediate reveal the flexibility of the active-site tunnel for catalysis. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **63**, 900–907.
- Thai, Q, Bös, F, and Pleiss, J (2009). The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. *BMC Genomics* **10** (1), 390.

- The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC, <http://pymol.org>* (2013).
- The UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **42** (D1), D191–D198.
- Thoma, R, Schulz-Gasch, T, D’Arcy, B, Benz, J, Aebi, J, Dehmlow, H, Hennig, M, Stihle, M, and Ruf, A (2004). Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* **432**, 118–122.
- Thompson, JD, Linard, B, Lecompte, O, and Poch, O (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One* **6**, e18093.
- Tittmann, K, Neef, H, Golbik, R, Hübner, G, and Kern, D (2005). Kinetic Control of Thiamin Diphosphate Activation in Enzymes Studied by Proton-Nitrogen Correlated NMR Spectroscopy. *Biochemistry* **44** (24), 8697–8700.
- Tittmann, K, Vyazmensky, M, Hübner, G, Barak, Z, and Chipman, DM (2005). The carboligation reaction of acetohydroxyacid synthase II: Steady-state intermediate distributions in wild type and mutants by NMR. *Proc Natl Acad Sci USA* **102** (3), 553–558.
- Tittmann, K, Wille, G, Golbik, R, Weidner, A, Ghisla, S, and Hübner, G (2005). Radical phosphate transfer mechanism for the thiamin diphosphate- and FAD-dependent pyruvate oxidase from *Lactobacillus plantarum*. Kinetic coupling of intercofactor electron transfer with phosphate transfer to acetyl-thiamin diphosphate via a transient FA. *Biochemistry* **44**, 13291–13303.
- Todd, AE, Orengo, CA, and Thornton, JM (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307**, 1113–1143.
- Triantafyllou, AO, Adlercreutz, P, and Mattiasson, B (1993). Influence of the reaction medium on enzyme activity in bio-organic synthesis: behaviour of lipase from *Candida rugosa* in the presence of polar additives. *Biotechnol Appl Biochem* **17**, 167–179.
- Tsou, AY, Ransom, SC, Gerlt, JA, Buechter, DD, Babbitt, PC, and Kenyon, GL (1990). Mandelate pathway of *Pseudomonas putida*: sequence relationships involving mandelate racemase, (*S*)-mandelate dehydrogenase, and benzoylformate decarboxylase and expression of benzoylformate decarboxylase in *Escherichia coli*. *Biochemistry* **29**, 9856–9862.
- Vogel, C and Pleiss, J (2014). The modular structure of ThDP-dependent enzymes. *Proteins* **82** (10), 2523–2537.

- Vogel, C, Widmann, M, Pohl, M, and Pleiss, J (2012). A standard numbering scheme for thiamine diphosphate-dependent decarboxylases. *BMC Biochem* **13** (1), 24.
- Vyazmensky, M, Steinmetz, A, Meyer, D, Golbik, R, Barak, Z, Tittmann, K, and Chipman, DM (2011). Significant Catalytic Roles for Glu47 and Gln110 in All Four of the C-C Bond-Making and -Breaking Steps of the Reactions of Acetohydroxyacid Synthase II. *Biochemistry* **50** (15), 3250–3260.
- Wagner, T, Bellinzoni, M, Wehenkel, A, O'Hare, HM, and Alzari, PM (2011). Functional plasticity and allosteric regulation of α -ketoglutarate decarboxylase in central mycobacterial metabolism. *Chem Biol* **18**, 1011–1020.
- Wang, J, Golbik, R, Seliger, B, Spinka, M, Tittmann, K, Hübner, G, and Jordan, F (2001). Consequences of a modified putative substrate-activation site on catalysis by yeast pyruvate decarboxylase. *Biochemistry* **40**, 1755–1763.
- Waterhouse, AM, Procter, JB, Martin, DMA, Clamp, M, and Barton, GJ (2009). Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25** (9), 1189–1191.
- Webb, B and Sali, A (2014). Protein Structure Modeling with MODELLER. *Protein Structure Prediction*. Ed. by D Kihara. **1137**. Methods in Molecular Biology. Springer New York, 1–15.
- Wechsler, Cindy (2014). Intermediates in the carbonylation of the ThDP-dependent pyruvate decarboxylase - Tailor-made enzyme catalysts. PhD thesis. Georg-August-University, Göttingen.
- Werther, T, Zimmer, A, Wille, G, Golbik, R, Weiss, MS, and König, S (2010). New insights into structure-function relationships of oxalyl CoA decarboxylase from *Escherichia coli*. *FEBS J* **277**, 2628–2640.
- Wescott, CR and Klibanov, AM (1994). The solvent dependence of enzyme specificity. *Biochim Biophys Acta* **1206**.1, 1–9.
- Westphal, R, Hahn, D, Mackfeld, U, Waltzer, S, Beigi, M, Widmann, M, Vogel, C, Pleiss, J, Müller, M, Rother, D, and Pohl, M (2013). Tailoring the *S*-selectivity of 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase (MenD) from *Escherichia coli*. *ChemCatChem* **5**, 3587–3594.
- Westphal, R, Jansen, S, Vogel, C, Pleiss, J, Müller, M, Rother, D, and Pohl, M (2014). MenD from *Bacillus subtilis*: a potent catalyst for the enantio-complementary asymmetric synthesis of functionalized α -hydroxy ketones. *ChemCatChem* **6**, 1082–1088.

- Westphal, R, Vogel, C, Schmitz, C, Pleiss, J, Müller, M, Pohl, M, and Rother, D (2014a). A Tailor-Made Chimeric Thiamine Diphosphate Dependent Enzyme for the Direct Asymmetric Synthesis of (*S*)-Benzoin. *Angew Chem Int Ed Engl* **53** (35), 9376–9379.
- (2014b). Ein maßgeschneidertes chimäres Thiamindiphosphat-abhängiges Enzym zur direkten asymmetrischen Synthese von (*S*)-Benzoinen. *Angew Chem Weinheim Bergstr Ger* **126** (35), 9530–9533.
- Westphal, R, Waltzer, S, Mackfeld, U, Widmann, M, Pleiss, J, Beigi, M, Müller, M, Rother, D, and Pohl, M (2013). (*S*)-Selective MenD variants from *Escherichia coli* provide access to new functionalized chiral α -hydroxy ketones. *Chem Commun (Camb)* **49**, 2061–2063.
- Westphal, Robert (2013). Tailor-made thiamine diphosphate-dependent enzymes for *S*-selective carbonylation - Complementation of the α -hydroxy ketone platform. PhD thesis. Institut für Bio- und Geowissenschaften (IBG-1: Biotechnologie) des Forschungszentrums Jülich GmbH.
- Widmann, M, Juhl, PB, and Pleiss, J (2010). Structural classification by the Lipase Engineering Database: a case study of *Candida antarctica* lipase A. *BMC Genomics* **11.1**, 123.
- Widmann, M, Pleiss, J, and Oelschlaeger, P (2012). Systematic Analysis of Metallo- β -Lactamases Using an Automated Database. *Antimicrob Agents Chemother* **56**, 3481–3491.
- Widmann, M, Radloff, R, and Pleiss, J (2010). The Thiamine diphosphate dependent Enzyme Engineering Database: A tool for the systematic analysis of sequence and structure relations. *BMC Biochem* **11**, 9.
- Wu, YG, Chang, AK, Nixon, PF, and Duggleby, RG (2000). Mutagenesis at Asp27 of pyruvate decarboxylase from *Zymomonas mobilis*. Effect on its ability to form acetoin and acetolactate. *Eur J Biochem* **267**, 6493–6500.
- Wynn, RM, Machius, M, Chuang, JL, Li, J, Tomchick, DR, and Chuang, DT (2003). Roles of His291- α and His146- β ' in the Reductive Acylation Reaction Catalyzed by Human Branched-chain α -Ketoacid Dehydrogenase: refined phosphorylation loop structure in the active site. *J Biol Chem* **278** (44), 43402–43410.
- Xiang, S, Usunow, G, Lange, G, Busch, M, and Tong, L (2007). Crystal structure of 1-deoxy-D-xylulose 5-phosphate synthase, a crucial enzyme for isoprenoids biosynthesis. *J Biol Chem* **282**, 2676–2682.
- Yan, Z, Fushinobu, S, and Wakagi, T (2014). Four Cys residues in heterodimeric 2-oxoacid:ferredoxin oxidoreductase are required for CoA-dependent oxidative decarboxylation but not for a non-oxidative decarboxylation. *Biochim Biophys Acta* **1844** (4), 736–743.

-
- Yep, A, Kenyon, GL, and McLeish, MJ (2006). Determinants of substrate specificity in KdcA, a thiamin diphosphate-dependent decarboxylase. *Bioorg Chem* **34** (6), 325–336.
- (2008). Saturation mutagenesis of putative catalytic residues of benzoylformate decarboxylase provides a challenge to the accepted mechanism. *Proc Natl Acad Sci USA* **105** (15), 5733–5738.
- Yep, A and McLeish, MJ (2009). Engineering the substrate binding site of benzoylformate decarboxylase. *Biochemistry* **48**, 8387–8395.
- Yi, J, Nemeria, N, McNally, A, Jordan, F, Machado, RS, and Guest, JR (1996). Effect of Substitutions in the Thiamin Diphosphate-Magnesium Fold on the Activation of the Pyruvate Dehydrogenase Complex from *Escherichia coli* by Cofactors and Substrate. *J Biol Chem* **271** (52), 33192–33200.
- Yoon, MY, Gedi, V, Kim, J, Park, Y, Kim, DE, Park, EH, and Choi, JD (2010). Structural and functional evaluation of three well-conserved serine residues in tobacco acetohydroxyacid synthase. *Biochimie* **92** (1), 65–70.
- Zbinden, G (1962). Therapeutic use of vitamin B1 in diseases other than beriberi. *Ann N Y Acad Sci* **98.2**, 550–561.
- Zhang, G, Dai, J, Lu, Z, and Dunaway-Mariano, D (2003). The phosphonopyruvate decarboxylase from *Bacteroides fragilis*. *J Biol Chem* **278**, 41302–41308.
- Zhang, S, Liu, M, Yan, Y, Zhang, Z, and Jordan, F (2004). C2- α -Lactylthiamin Diphosphate Is an Intermediate on the Pathway of Thiamin Diphosphate-dependent Pyruvate Decarboxylation: Evidence on enzymes and models. *J Biol Chem* **279** (52), 54312–54318.

A Supporting Information

A.1 Structural rearrangement of *ApPDC*

The amino acid and DNA sequences of the rearranged *ApPDC* variant with a N-terminal PP and a C-terminal PYR domain (Section 2.3.2):

```
M A G L T N D E I V R H I N A L L T S N T T L V A 25
ATGGCCGGTCTGACGAATGACGAAATCGTCCGTCAATCAACGCCCTGCTGACATCAAACACGACGCTGGTGGCA 75
E T G D S W F N A M R M T L P R G A R V E L E M Q 50
GAAACCGGCGATTTCATGGTTCAATGCCATGCGCATGACCCTGCCGCGCGGTGCGCGCGTGGAACTGAAAATGCAG 150
W G H I G W S V P S A F G N A M G S Q D R Q H V V 75
TGGGGCCATATCGGCTGGTCCGTGCCCTCCGCCTTCGGCAATGCCATGGGCTCGCAGGACCGCCAGCATGTGGTG 225
M V G D G S F Q L T A Q E V A Q M V R Y E L P V I 100
ATGGTAGGCGATGGCTCCTTCCAGCTTACCGCGCAGGAAGTGGCTCAGATGGTGCCTACGAACTGCCCGTCATT 300
I F L I N N R G Y V I E I A I H D G P Y N Y I K N 125
ATCTTTCTGATCAACAACCGTGGCTATGTCATTGAAATCGCCATTTCATGACGGCCCGTACAACCTATATCAAGAAC 375
W D Y A G L M E V F N A G E G H G L G L K A T T P 150
TGGGATTACGCCGGCTGATGGAAGTCTTCAACGCCGGAGAAGGCCATGGACTTGGCCTGAAAGCCACCACCCCG 450
K E L T E A I A R A K A N T R G P T L I E C Q I D 175
AAGAACTGACAGAAGCCATCGCCAGGGCAAAAGCCAATACCCGCGGCCCCGACGCTGATCGAATGCCAGATCGAC 525
R T D C T D M L V Q W K V A S T N A S G T T L A L 200
CGCACGGACTGCACGGATATGCTGGTTCAATGGAAGTTGCCTCAACCAACGCGTCAGGCACCACTCTGGCCCTC 600
E V G M Y L A E R L V Q I G L K H H F A V A G D Y 225
GAGTTGGCATGTATCTTGCAGAACGCCTTGTACAGATCGGGTGAAGCATCACTTCGCCGTGGCGGGCGACTAC 675
N L V L L D Q L L L N K D M K Q I Y C C N E L N C 250
AATCTCGTTCTTCTGGATCAGTTGCTCCTCAACAAGGACATGAAACAGATCTATTGCTGCAATGAGTTGAACTGT 750
G F S A E G Y A R S N G A A A A V V T F S V G A I 275
GGCTTCAGCGCGGAAGGCTACGCCCGTTCTAACGGGGTGCAGCAGCGTTGTCACCTTCAGCGTTGGCGCCATT 825
S A M N A L G G A Y A E N L P V I L I S G A P N S 300
TCCGCCATGAACGCCCTCGGCGGCGCCTATGCCGAAAACCTGCCGGTTATCCTGATTTCCGGCGCGCCCAACAGC 900
N D Q G T G H I L H H T I G K T D Y S Y Q L E M A 325
AATGATCAGGGCACAGGTCATATCCTGCATCACACAATCGGCAAGACGGATTACAGCTACCAGCTTAAAATGGCC 975
R Q V T C A A E S I T D A H S A P A K I D H V I R 350
CGTCAGGTCACCTGTGCCCGGAAAGCATTACCGACGCTCACTCCGCCCCGGCCAAGATTGACCACGTCATTTCG 1050
T A L R E R K P A Y L D I A C N I A S E P I E V 374
ACGGCGCTGCGGAGCGTAAGCCGGCCTATCTGGACATCGCGTGCAACATTGCCTCCGAGCCCATCGAAGTG 1122
```

A.2 *SERgrid*

SERgrid was designed as a method for the identification of positions in different protein structures that have the same functional role. While the standard numbering schemes presented in this thesis are limited to positions with equivalent backbone positions, the structure based approach of *SERgrid* was intended to additionally identify positions with deviating backbone positions but shared interactions. A simplified model was used to calculate the location of 'effect points' of amino acids: the functional moieties of all amino acids were reduced to the 'effective atom' that exhibits the effect and the 'neighbor atom', which is adjacent to the former mentioned atom. The 'effect point' was calculated by elongation of the axis connecting the 'neighbor' and 'effective' atoms by 2 Å in case of hydrogen donors and 2.6 Å in case of hydrogen acceptors (Figure A.1).

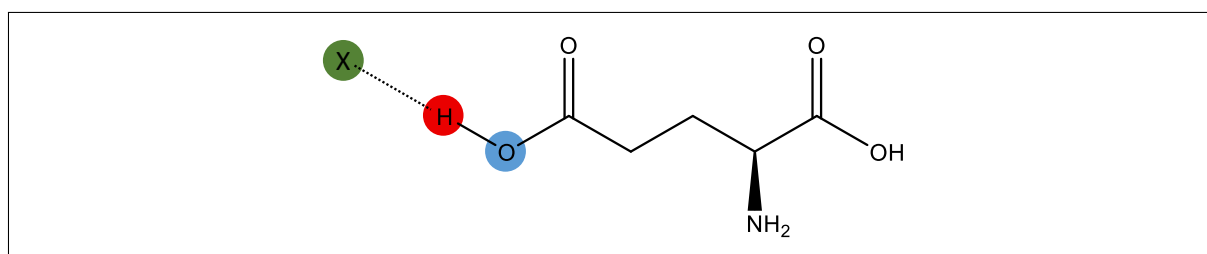


Figure A.1: The simplified model of the identification of 'effect points' (X, green) in enzyme structures exemplarily shown for the H-bond donor property of glutamic acid. By elongation of the axis connecting the 'neighbor atom' (blue) and the 'effective atom' (ref), the three-dimensional coordinates of a virtual 'effect point' (green) are defined.

By applying this model to a library of amino acid rotamers (Shetty et al. 2003), an extended library containing the 3D-coordinates of the atoms of all amino acids and of the respective 'effect points' was generated. Subsequently, the available protein structures of ThDP-dependent decarboxylases were superimposed using STAMP (Russell and Barton 1992) and shifted on all three axis to transform the aligned structures into the positive octant of the euclidean coordinate system. The three-dimensional space filled with the protein structures was further sub-divided into smaller boxes with a defined edge length, resulting in a 'grid'. Subsequently, each protein structure was individually processed in order to assign the possible location of all effect points to the boxes in the three-dimensional grid. Therefore, each amino acid of the protein structure was substituted by the respective rotamers encompassed in the extended library by mapping the backbone atoms of the rotamers onto the respective atoms in the crystal structure. Rotamers clashing the backbone of the protein were discarded¹. For those boxes of the grid containing

¹A clash was defined as a distance between any non-backbone atom of the rotamer and any atom of the protein backbone of less than 0.8 Å.

'effect points', the information was stored, which residues from which proteins might have an effect on the respective space. Subsequently, by taking the information of all superimposed structures together, the boxes were analyzed for shared interactions of residues from different proteins. In order to analyze the influence of the box size and the location of the grid, both parameters were varied. The analysis of *Pp*BFD, *Pf*BAL and *Ap*PDC resulted, depending on the box size (0.5 - 2.0 Å), in 4201 to 8624 pairs of positions with different backbone location but putatively comparable interaction with ligands like cofactors, substrates, products or other amino acids. Analysis of different shift intervals of the grid revealed minor influence of the exact position of the grid on the result but variation of the box size from 0.5 - 2.0 Å doubled the amount of identified residues that putatively have the same effect on nearby positions in the three-dimensional space.

A.3 Variants of ThDP-dependent enzymes

An extensive investigation of the literature on ThDP-dependent enzymes revealed 97 individual positions in the PYR and PP domains of ThDP-dependent enzymes to have been mutated in at least one enzyme (Table A.1). The mutated positions were addressed with their standard numbers using the domain based standard numbering scheme for ThDP-dependent enzymes (see Section 2.2.2). For organism abbreviations and references see specifications on pages 185ff.

Table A.1: Variants of ThDP-dependent enzymes identified in the literature.

continued Table A.1

std. pos.	variants	std. pos.	variants
26	<i>Pp</i> BFDC-P24A ^[1]		<i>Ec</i> AHASI-E60A/Q ^[28]
27	<i>Pp</i> BFDC-G25A ^[2]		<i>Ec</i> AHASII-E47A/Q ^[28,29,30]
28	<i>Az</i> CDH-H28A ^[3]		<i>Ec</i> DXPS-E370A ^[31]
	<i>Ec</i> IIPDC-D29E ^[4]		<i>Ec</i> GXC-V51D/E/S ^[32,33]
	<i>Ec</i> MenD-S32A/C/D/F/G/H/I/L /N/R/V/Y ^[5,6]		<i>Ec</i> MenD-E55D/Q ^[6,34]
	<i>Pf</i> BAL-A28S ^[7,8,9,10]		<i>Ec</i> PDHE1-E571A/D/Q ^[35]
	<i>Pp</i> BFDC-S26A/L/M/T ^[7,9,11,12]		<i>Ef</i> AHAS-E49A/D/Q ^[36]
	<i>Sc</i> PDC-D28A/N ^[13,14,15,16,17,18]		<i>Hs</i> PDH-E89A/D/Q ^[37]
	<i>Sv</i> PPDC-S25D/N ^[19]		<i>Nt</i> AHASI-E143A ^[27]
	<i>Zm</i> PDC-D27A/E/G/N/S ^[20,21,22,23]		<i>Of</i> OCDC-E56A ^[38]
29	<i>Bs</i> MenD-R32A ^[24]		<i>Sc</i> PDC-E51A/D/Q/X ^[17,18,39,40]
	<i>Ec</i> AHASII-A26V ^[25]		<i>Sc</i> TK-E418A ^[41]
	<i>Ec</i> MenD-R33K ^[6]	52	<i>Ec</i> GXC-E52Q ^[32]
	<i>Pf</i> BAL-H29A ^[7]	56	<i>Ec</i> PDHE1-C576A ^[45]
49	<i>Nt</i> AHASI-R141A/F/K ^[26]	69	<i>Pp</i> BFDC-F66I ^[46]
50	<i>Nt</i> AHASI-H142K/T ^[27]	73	<i>Ec</i> AHASI-C83A/S/T ^[28]
51	<i>Ec</i> IIPDC-E52D ^[4]		<i>Ec</i> DXPS-Y392A/F ^[31]
			<i>Ec</i> TK-F434A ^[47]
			<i>Gs</i> PDH-Q81E ^[48]

continued Table A.1

std. pos.	variants
	<i>Pp</i> BFDC-H70A/F/L/Q/S/T [7,11,12,49]
74	<i>Nt</i> AHASI-S167A/F/R [50]
76	<i>Mt</i> AHAS-P111A/E/T/V [51]
80	<i>Ec</i> DXPS-R398A [31] <i>Ec</i> PDHE1-R606A [35]
85	<i>Ec</i> PDHE1-C611A [45]
86	<i>Rn</i> TK-R433A [52]
89	<i>Pp</i> BFDC-W86R [46]
91	<i>Sc</i> PDC-E91A/D/Q [17,18,53,54]
92	<i>Suc</i> A-R710A [55] <i>Sc</i> PDC-H92A/C/G/K [18,56]
101	<i>Ec</i> TK-H461Q/S/Y [57,58] <i>Sc</i> TK-H469N/Q/S [59]
102	<i>Bs</i> MenD-R106A [24] <i>Ec</i> AHASII-V99M [25] <i>Ec</i> MenD-R107K [6]
111	<i>Sc</i> PDC-L111A/V/W [60]
112	<i>Pp</i> BFDC-E107R [61] <i>Zm</i> PDC-V111A [62]
113	<i>Ec</i> AHASII-A108V [25] <i>Zm</i> PDC-L112A [63]
114	<i>EC</i> IPDC-H115K [4] <i>Ec</i> AHASII-F109M [64] <i>Ef</i> AHAS-H111F/R [65] <i>Of</i> OCDC-Y120A/F [38] <i>Pp</i> BFDC-L109A/X [61,66] <i>Sc</i> PDC-H114F [14,18] <i>Sv</i> PPDC-H110A [19]

continued Table A.1

std. pos.	variants
	<i>Zm</i> PDC-H113A/K/Q/R [23,62,67]
115	<i>Ec</i> AHASII-Q110A/E/H/N [30] <i>Ec</i> DXPS-H431A [31] <i>Ef</i> AHAS-Q112E/N/V [65] <i>Gs</i> PDH-H128A/N/Q [48,68] <i>Of</i> OCDC-E121A/Q [38] <i>Pf</i> BAL-Q113A/H [7] <i>Pp</i> BFDC-L110A/X [2,61,66] <i>Sc</i> PDC-H115F [14,18] <i>Sc</i> TK-H481A [41] <i>Sv</i> PPDC-H111A [19] <i>Zm</i> PDC-H114A/Q [23,62,67]
119	<i>Sc</i> PDC-N119K [40]
120	<i>Zm</i> PDC-T119I [69]
124	<i>Pp</i> BFDC-D114R [61]
130	<i>Pp</i> BFDC-R120E [61]
133	<i>Ec</i> PDHE1-C655A [45] <i>Zm</i> PDC-T132A [69]
151	<i>Pp</i> BFDC-R141E [61]
152	<i>Sc</i> PDC-C152A [18,56]
168	<i>Sc</i> TK-S527A/C/G/P/R [59]
169	<i>Ec</i> AHASII-K159A/M/Q [30] <i>Ec</i> DXPS-R478A [31] <i>Ec</i> TK-R520G/I/P/Q/V [57,58] <i>Rn</i> TK-R506A [52] <i>Sc</i> TK-R528K/N/Q/T/Y [59]
170	<i>Sc</i> TK-Q529I/K/N/R/S/T [59]
180	<i>Nt</i> AHASI-W266F [70]

continued Table A.1

std. pos.	variants
188	<i>Pp</i> BFDC-S181T [46]
221	<i>Sc</i> PDC-C221A/E/S [17,18,71,72,73]
222	<i>Sc</i> PDC-C222A/S [17,18,72,73]
264	<i>Ec</i> AHASII-G249A/E/V [25]
266	<i>Ec</i> AHASI-M263A [28] <i>Ec</i> AHASII-M250A [74,75]
267	<i>Ec</i> AHASII-H251A/N/Q [30]
288	<i>Nt</i> AHASI-R372F/K [26]
291	<i>Sc</i> PDC-D291A/N [40,60]
292	KdcA-S286Y [76] <i>Ec</i> AHASI-R289K/Q [28] <i>Ec</i> AHASII-R276K/Q [30,75] <i>Ec</i> MenD-K292Q [6] <i>Nt</i> AHASI-R376F/K [26] <i>Pp</i> BFDC-H281A/F/N/Q/T/W/X/Y [11,12,66,77] <i>Sc</i> PhePDC-I335Y [78]
293	<i>Ec</i> MenD-R293K [6] <i>Sc</i> PDC-N293A [60]
294	<i>Bs</i> MenD-K299A [24] <i>Sc</i> PDC-T294A [60]
296	<i>Pf</i> BAL-H286A [10]
298	<i>Sc</i> PDC-S298A [60]
300	<i>Pp</i> BFDC-Y288A [61] <i>Sc</i> PDC-S300A [60]
313.4	<i>Pp</i> BFDC-A306F [61]
346.1	<i>Nt</i> AHASI-W439F [70]
364	<i>Sc</i> PDC-P364H [40]
374	<i>Ec</i> PDHE1-C120A [45]

continued Table A.1

std. pos.	variants
377	<i>Pp</i> BFDC-M365L [79]
383	<i>Rn</i> TK-R102A [52]
388	<i>Ap</i> PDC-T384G [80] <i>Ec</i> IIIPDC-Q383T [4] <i>Ec</i> AHASI-V391A [28] <i>Ec</i> AHASII-V375A/I [64] <i>Ec</i> GXC-I393A [33]
389	<i>Ec</i> MenD-S391A [6] <i>Pp</i> BFDC-T377L [66]
390	<i>St</i> KGOR-C46A [81]
392	KdcA-F381W [76] <i>Ap</i> PDC-W388A/I [82,83] <i>Pp</i> BFDC-T380X [66] <i>Sc</i> PhePDC-Q448W [78] <i>Zm</i> PDC-W392A/E/F/G/H/I/M/ N/Q/V [84,85]
393	<i>Bs</i> MenD-R409A [24] <i>Ec</i> MenD-R395A/K/Y [5,6] <i>Nt</i> AHASI-M489V [27]
394	<i>Bf</i> PPDC-E213A [86] <i>Nt</i> AHASI-W490F [70] <i>Sv</i> PPDC-E224A [19]
398	<i>Nt</i> AHASI-Y494H [27]
406	<i>Nt</i> AHASI-W503F [70]
408	<i>Pp</i> BFDC-F397A/X [2,66]
409	<i>Nt</i> AHASI-S506A/F/R [50] <i>Pp</i> BFDC-C398X [66]
410	<i>Bs</i> MenD-R428A [24] <i>Ec</i> MenD-R413K [6]

continued Table A.1

std. pos.	variants
412	<i>Sc</i> PDC-W412A/F [87]
413	<i>Ef</i> AHAS-Q411E/N [65] <i>Zm</i> PDC-G413A/V [20]
414	<i>Pp</i> BFDC-G401X [66]
415	<i>Ec</i> MenD-I418L [6] <i>Pp</i> BFDC-L403X [66] <i>Sc</i> PDC-I415A/C/L/M/S/T/V [18,88]
429	<i>Suc</i> A-R337A [55]
442	<i>Bf</i> PPDC-D258A [86] <i>Sv</i> PPDC-D263A [19]
444	<i>Bf</i> PPDC-D260A [86] <i>Ec</i> AHASII-D428E/N [29] <i>Hs</i> 2HCL-D455R/S [89] <i>Zm</i> PDC-D440E/G/N/T [62,90]
445	<i>Ec</i> PDHE1-G231A/S [18,91]
446	<i>Nt</i> AHASI-S539A/F/R [50] <i>Rs</i> TK-C160A/D/E/S [92]
449	<i>Ec</i> PDHE1-E235A [35] <i>Nt</i> AHASI-M542C/I/V [27]
450	<i>Ec</i> PDHE1-E237A [35] <i>Gs</i> PDH-D180A/N [93]
453	<i>Gs</i> PDH-E183A/Q [93] <i>Zm</i> PDC-E449D/N [62,69]
455	<i>Sc</i> PDC-S455F [40]
463	<i>Zm</i> PDC-P459A/G [69]
469	<i>Ec</i> PDHE1-N258Q [18,91]
470	<i>Ec</i> PDHE1-C259N/S [18,91]
471	<i>Zm</i> PDC-N467D/Q [62]

continued Table A.1

std. pos.	variants
473	<i>Ec</i> TK-S188Q/R/T [57]
475	<i>Bs</i> MenD-G488Q [94] <i>Ec</i> AHASII-G459A [95] <i>Of</i> OCDC-Y483A/F [38] <i>Sv</i> PPDC-D297E/N [19]
476	<i>Kdc</i> A-V461I [76] <i>Ap</i> PDC-I468A/G/V [80] <i>Bs</i> MenD-I489A/G [24,94] <i>Ec</i> AHASI-L476M [28] <i>Ec</i> AHASII-M460A [95] <i>Ec</i> GXC-L478A [33] <i>Ec</i> MenD-I474A/G [96] <i>Pf</i> BAL-A480I [10] <i>Pp</i> BFDC-A460F/G/I/L/Y [1,66,97,98,99] <i>Zm</i> PDC-I472A/F/L/S [63,97]
477	<i>Ap</i> PDC-E469G [80,82] <i>Az</i> CDH-N484A [3] <i>Bs</i> MenD-F490A/G [24,94] <i>EC</i> IIPDC-E468D [4] <i>Ec</i> AHASI-V477I [28] <i>Ec</i> AHASII-V461G [95] <i>Ec</i> GXC-I479V [33] <i>Ec</i> MenD-F475A/G [96] <i>Pp</i> BFDC-L461A/G/S/V/X [1,66,79] <i>Sc</i> PDC-E477D/N/Q [13,14,15,16,17,18,100] <i>Sc</i> PhePDC-E545L [78] <i>Zm</i> PDC-E473A/C/D/F/G/H/I/ L/N/Q/S/T/V [20,22,23,42,63,101]

continued Table A.1

std. pos.	variants
480	<i>AzCDH</i> -L487A ^[3] <i>BsMenD</i> -L493A ^[24] <i>EcAHASI</i> -Q480W ^[28] <i>EcAHASII</i> -W464L ^[74,75] <i>EcMenD</i> -L478D/F/G/H/I/N/R/ S/V/Y ^[5] <i>LpPOX</i> -E483A/Q ^[9] <i>NtAHASI</i> -W573F ^[70] <i>PfBAL</i> -F484I ^[10] <i>PpBFDC</i> -F464I/X ^[66,97,98,99] <i>ZmPDC</i> -I476A/E/F/L/V ^[63,97]
488	<i>PpBFDC</i> -L476A/C/G/H/K/M/P/ Q/S/T ^[46,79] <i>ScTK</i> -I191A/L/V ^[59] <i>ZmPDC</i> -N482A/D/S ^[63]
493	<i>ZmPDC</i> -W487L ^[90]
502	<i>ZmPDC</i> -F496H/I/L ^[43]
541	<i>GsPDH</i> -F266A ^[102] <i>PpBFDC</i> -S525G ^[46]
542	<i>GsPDH</i> -R267A ^[102] <i>StKGOR</i> -C197A ^[81]
552	<i>KdcA</i> -M538W ^[76] <i>ApPDC</i> -W543F ^[80]
559	<i>ScPhePDC</i> -M624W ^[78]

Besides those variants (Table A.1), the referenced literature contains information about variants outside the PYR and PP domains and variants in regions of the PYR and PP domains deviating considerably between different superfamilies. Thus, the variants mentioned below (Table A.2) can not be compared to members of all other superfamilies. For organism abbreviations and references see specifications on pages 185ff.

Table A.2: Variants mutated at positions outside of the PYR and PP domains of ThDP-dependent enzymes identified in the literature.

Variant	Information of this variant can not be transferred to other superfamilies using the standard numbering scheme, since the mutated position ...	Ref.
<i>HsBcaKADH</i> -H146A/N	forms part of the structurally variable region that corresponds to the α -helix PYR- α E of decarboxylases	[103]
<i>HsBcaKADH</i> -H291A/Q/N	is located on the structurally variable N-terminus of the PP domain	[103]
<i>EcPDHE1</i> - Y177A/F	forms part of a large insertion between PP-B and PP-C (referred to as 'region 1' in section 4.2.4 on pages 104ff. and figure 4.5 on page 105)	[104]
<i>EcPDHE1</i> - E401K	is located on the linker between the PP and PYR domains	[105]
<i>EcPDHE1</i> - H407A	is located on the linker between the PP and PYR domains	[106]
<i>EcPDHE1</i> - E636A/Q	forms part of the structurally variable region that corresponds to the α -helix PYR- α E of decarboxylases	[15,107]
<i>EcPDHE1</i> - C771S	is located on the linker between the PYR and TKC domains	[45]
<i>GsPDHE1</i> - I206A	forms part of the TKC domain	[48]
<i>GsPDHE1</i> - H271A	is located on the structurally variable N-terminus of the PP domain	[68]
<i>GsPDHE1</i> - D276A	is located on the structurally variable N-terminus of the PP domain	[102]

continued Table A.2

Variant	Information of this variant can not be transferred to other superfamilies using the standard numbering scheme, since the mutated position ...	Ref.
<i>Gs</i> PDHE1- S283A	is located on the structurally variable N-terminus of the PP domain	[102]
<i>SucA</i> - H313	forms part of the insert in front of PP- β 2 of aKGDHs	[108]
<i>SucA</i> - H460I	is located on the linker between the PP and PYR domains	[109]
<i>SucA</i> - W533A	is located on the linker between the PP and PYR domains	[110]
<i>Ms</i> OGDC- E1034A	forms part of the ATN domain	[110]
<i>Ms</i> OGDC- R1062A	is located on an elongated loop between PYR- α F and PYR- β 5	[110]
<i>St</i> KGOR- C12A	corresponds to the linker between the TH3 and PP domains of decarboxylases	[81]
<i>St</i> KGOR- C15A	corresponds to the linker between the TH3 and PP domains of decarboxylases	[81]
<i>Rn</i> TK- R350A	forms part of PYR- α A, which is connected to the linker from the PP domain and is slightly shifted as compared to decarboxylases. Thus, this position does not perfectly correspond to a position in decarboxylases. This position corresponds to position 358 in <i>Ec</i> TK.	[52]
<i>Sc</i> TK- H103A	forms part of the insert between PP- α B and PP- β 2. This position corresponds to position 100 in <i>Ec</i> TK.	[111]
<i>Sc</i> TK- R359A/I/P/T	forms part of PYR- α A, which is connected to the linker from the PP domain and is slightly shifted as compared to decarboxylases. Thus, this position does not perfectly correspond to a position in decarboxylases. This position corresponds to position 358 in <i>Ec</i> TK.	[59]

continued Table A.2

Variant	Information of this variant can not be transferred to other superfamilies using the standard numbering scheme, since the mutated position ...	Ref.
<i>EcTK</i> - H26A/K/T/V/Y	is located at the N-terminus. Thus, it corresponds to a residue forming the linker between the TH3 and PP domains of decarboxylases	[57,112,113]
<i>EcTK</i> - A29D/E	is located at the N-terminus. Thus, it corresponds to a residue forming the linker between the TH3 and PP domains of decarboxylases	[57]
<i>EcTK</i> - H100A/I/V	forms part of the insert between PP- α B and PP- β 2. This position corresponds to position 103 in <i>ScTK</i> .	[57]
<i>EcTK</i> - D259A/G	is located on the linker between the PP and PYR domains	[57]
<i>EcTK</i> - D261A	is located on the linker between the PP and PYR domains	[112]
<i>EcTK</i> - R358I/L/P	forms part of PYR- α A, which is connected to the linker from the PP domain and is slightly shifted as compared to decarboxylases. Thus, this position does not perfectly correspond to a position in decarboxylases. This position corresponds to position 350 in <i>RnTK</i> .	[57,58]
<i>EcTK</i> - D469A/E/K/S/T/Y	forms part of the region between PYR- β 4 and PYR- α F, which deviates between the different superfamilies	[57,113]
<i>EcDXPS</i> - H49Q	forms part of a 3-helix domain at the N-terminus of <i>EcDXPS</i> . Such a domain does not exist in other superfamilies.	[114]
<i>LpPOX</i> - T561A	is located at the C-terminus that differs between structures of different decarboxylases	[9]
<i>LpPOX</i> - S562A	is located at the C-terminus that differs between structures of different decarboxylases	[9]
<i>OfOCD</i> - S553A	is located at the N-terminus that differs between structures of different decarboxylases	[38]

continued Table A.2

Variant	Information of this variant can not be transferred to other superfamilies using the standard numbering scheme, since the mutated position ...	Ref.
<i>Of</i> OCDC- R555A	is located at the N-terminus that differs between structures of different decarboxylases	[38]

Protein names abbreviated in tables A.1 and and A.2 (the respective superfamilies are declared in paranthesis):

*Ap*PDc, *Acetobacter pasteurianus* Pyruvate decarboxylase (DC); *Az*CDH, *Azoarcus sp.* Cyclohexane-1,2-dione hydrolase (DC); *Bf*PPDC, *Bacteroides fragilis* Phosphonopyruvate decarboxylase (PPDC); *Bs*MenD, *Bacillus subtilis* 2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase (DC); *Ec*AHASI, *Escherichia coli* Acetohydroxy-acid synthase I catalytic subunit (DC); *Ec*AHASII, *E. coli* Acetohydroxy-acid synthase II catalytic subunit (DC); *Ec*DXPS, *E. coli* 1-deoxy-D-xylulose-5-phosphate synthase (DXPS); *Ec*GXC, *E. coli* Glyoxylate carboligase (DC); *EC*IPDC, *Enterobacter cloacae* Indole-3-pyruvate decarboxylase (DC); *Ec*MenD, *E. coli* 2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate synthase (DC); *Ec*PDHE1, *E. coli* Pyruvate dehydrogenase E1 component (aKADH1); *Ec*TK, *E. coli* Transketolase (TK); *Ef*AHAS, *Enterococcus faecalis* V583 Acetohydroxy-acid synthase (DC); *Gs*PDH, *Geobacillus stearothermophilus* Pyruvate dehydrogenase E1 component beta subunit (aKADH2); *Hs*BcaKADH, human branched-chain α -Ketoacid dehydrogenase (aKADH1); *Hs*HCL, human 2-Hydroxyacyl-CoA lyase 1 (DC); *Hs*PDH, human Pyruvate dehydrogenase E1 component beta subunit (aKADH2); *KdcA*, *Lactococcus lactis* branched-chain α -Ketoacid decarboxylase (DC); *Lp*POX, *Lactobacillus plantarum* Pyruvate oxidase (DC); *Ms*OGDC, *Mycobacterium smegmatis* α -Ketoglutarate decarboxylase (aKGDH); *Mt*AHAS, *Mycobacterium tuberculosis* Acetohydroxy-acid synthase I catalytic subunit (DC); *Nt*AHASI, *Nicotiana tabacum* Acetohydroxy-acid synthase I catalytic subunit (DC); *Of*OCDC, *Oxalobacter formigenes* Oxalyl-CoA decarboxylase (DC); *Pf*BAL, *Pseudomonas fluorescens* Benzaldehyde lyase (DC); *Pp*BFDC, *Pseudomonas putida* Benzoylformate decarboxylase (DC); *Rn*TK, *Rattus norvegicus* Transketolase (TK); *Rs*TK, *Rhodobacter sphaeroides* Transketolase (TK); *Sc*PDc, *Saccharomyces cerevisiae* Pyruvate decarboxylase (DC); *Sc*PhePDc, *S. cerevisiae* Phenylpyru-

vate decarboxylase (DC); *ScTK*, *S. cerevisiae* Transketolase (TK); *StKGOR*, *Sulfolobus tokodaii* 2-Oxoglutarate ferredoxin, oxidoreductase beta subunit (OR); *SucA*, *E. coli* 2-Oxoglutarate dehydrogenase E1 component (aKGDH); *SvPPDC*, *Streptomyces viridochromogenes* Phosphonopyruvate decarboxylase (PPDC); *ZmPDC*, *Zymomonas mobilis* Pyruvate decarboxylase (DC)

References used in tables A.1 and A.2:

[¹]Gocke, Walter, et al. 2008, [²]Siegert 2000, [³]Loschonsky 2014, [⁴]Schütz, Golbik, et al. 2005, [⁵]Westphal, Hahn, et al. 2013, [⁶]Fang, Macova, et al. 2011, [⁷]Kneen, Pogozeva, et al. 2005, [⁸]Brandt, Kneen, et al. 2010, [⁹]Meyer 2009, [¹⁰]Janzen et al. 2006, [¹¹]Polovnikova et al. 2003, [¹²]Yep, Kenyon, and McLeish 2008, [¹³]Kutter et al. 2009, [¹⁴]Liu et al. 2001, [¹⁵]Baykal et al. 2006, [¹⁶]Sergienko and Jordan 2001b, [¹⁷]Balakrishnan, Gao, et al. 2012, [¹⁸]Jordan, Nemeria, et al. 1998, [¹⁹]Johnen and Sprenger 2009, [²⁰]Wechsler 2014, [²¹]Wu et al. 2000, [²²]Chang, Nixon, and Duggleby 1999, [²³]Huang et al. 2001, [²⁴]Dawson et al. 2010, [²⁵]Hill and Duggleby 1998, [²⁶]Le et al. 2005, [²⁷]Le et al. 2004, [²⁸]Belenky et al. 2012, [²⁹]Bar-Ilan et al. 2001, [³⁰]Vyazmensky et al. 2011, [³¹]Xiang et al. 2007, [³²]Kaplan et al. 2008, [³³]Nemeria, Binshtein, et al. 2012, [³⁴]Bhasin, Billinsky, and Palmer 2003, [³⁵]Nemeria, Arjunan, et al. 2010, [³⁶]Lee, Lee, et al. 2013, [³⁷]Fang, Nixon, and Duggleby 1998, [³⁸]Berthold, Toyota, et al. 2007, [³⁹]Killenberg-Jabs et al. 1997, [⁴⁰]Eberhardt et al. 1999, [⁴¹]Kern et al. 1997, [⁴²]Lie et al. 2005, [⁴³]Candy, Koga, et al. 1996, [⁴⁴]Tittmann, Neef, et al. 2005, [⁴⁵]Nemeria, Volkov, et al. 1998, [⁴⁶]Lingen, Grötzinger, et al. 2002, [⁴⁷]Galman et al. 2010, [⁴⁸]Pei et al. 2008, [⁴⁹]Sergienko, Wang, et al. 2000, [⁵⁰]Yoon et al. 2010, [⁵¹]Baig et al. 2013, [⁵²]Soh et al. 1998, [⁵³]Li, Furey, and Jordan 1999, [⁵⁴]Zhang, Liu, et al. 2004, [⁵⁵]Frank, Leeper, and Luisi 2007, [⁵⁶]Baburina, Li, et al. 1998, [⁵⁷]Hibbert et al. 2008, [⁵⁸]Payongsri et al. 2012, [⁵⁹]Ranoux et al. 2012, [⁶⁰]Joseph et al. 2006, [⁶¹]Andrews, Rogers, et al. 2014, [⁶²]Candy and Duggleby 1998, [⁶³]Pohl, Siegert, et al. 1998, [⁶⁴]Steinmetz et al. 2010, [⁶⁵]Lee, Jung, et al. 2015, [⁶⁶]Yep and McLeish 2009, [⁶⁷]Schenk, Leeper, et al. 1997, [⁶⁸]Fries, Jung, and Perham 2003, [⁶⁹]Candy and Duggleby 1994, [⁷⁰]Chong et al. 1999, [⁷¹]Lu et al. 2000, [⁷²]Wang et al. 2001, [⁷³]Baburina, Gao, et al. 1994, [⁷⁴]Engel et al. 2003, [⁷⁵]Tittmann, Vyazmensky, et al. 2005, [⁷⁶]Yep, Kenyon, and McLeish 2006, [⁷⁷]Dünkelmann et al. 2002, [⁷⁸]Kneen, Stan, et al. 2011, [⁷⁹]Lingen, Kolter-Jung, et al. 2003, [⁸⁰]Westphal et al. 2014a; Westphal et al. 2014b, [⁸¹]Yan, Fushinobu, and Wakagi 2014, [⁸²]Rother et al. 2011, [⁸³]Westphal 2013, [⁸⁴]Iding, Siegert, and Pohl 1998, [⁸⁵]Bruhn et al. 1995, [⁸⁶]Zhang, Dai, et al. 2003, [⁸⁷]Li and Jordan 1999, [⁸⁸]Guo et al. 1998, [⁸⁹]Fraccascia et al. 2011, [⁹⁰]Diefenbach et al. 1992, [⁹¹]Yi et al. 1996, [⁹²]Bobst and Tabita 2004, [⁹³]Frank, Titman, et al. 2004, [⁹⁴]Westphal, Jansen, et al.

2014, ^[95]Schmitz 2012, ^[96]Westphal, Waltzer, et al. 2013, ^[97]Siegert et al. 2005, ^[98]Kara et al. 2011, ^[99]Berheide et al. 2010, ^[100]Jordan, Zhang, and Sergienko 2002, ^[101]Meyer, Walter, et al. 2011, ^[102]Fries, Chauhan, et al. 2003, ^[103]Wynn et al. 2003, ^[104]Nemeria, Yan, et al. 2001, ^[105]Balakrishnan, Nemeria, et al. 2012, ^[106]Arjunan et al. 2006, ^[107]Nemeria, Tittmann, et al. 2005, ^[108]Frank, Price, et al. 2007, ^[109]Beigi et al. 2013, ^[110]Wagner et al. 2011, ^[111]Selivanov et al. 2004, ^[112]Asztalos et al. 2007, ^[113]Smith et al. 2008, ^[114]Querol et al. 2001

A.4 A standard numbering scheme for thiamine diphosphate-dependent decarboxylases

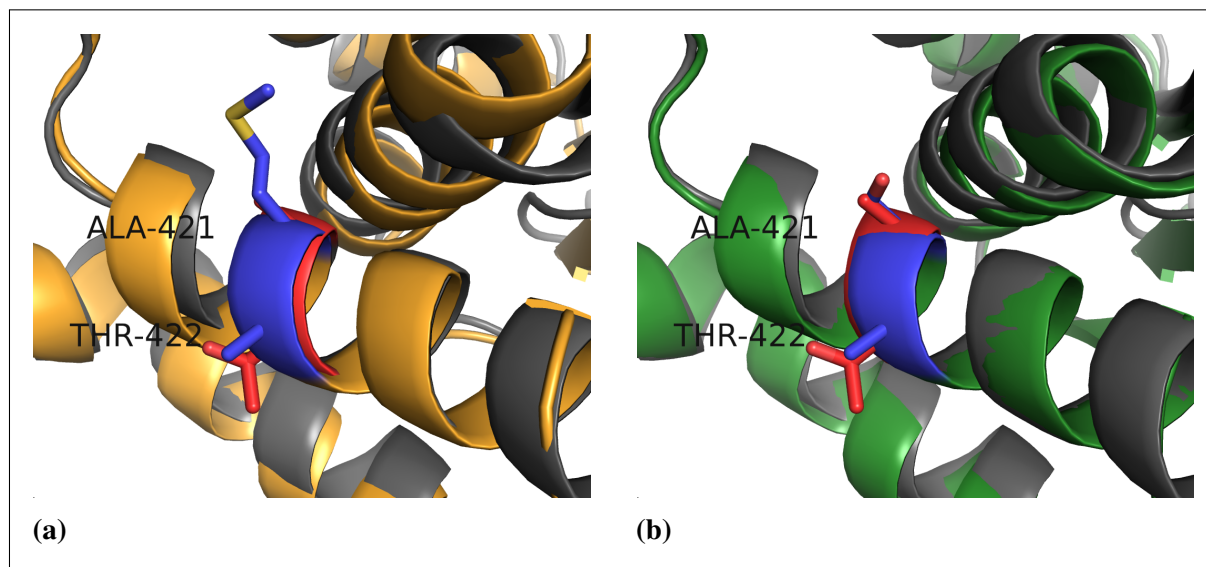


Figure A.2: Divergent positions 421 and 422 in the comparison of the alignment methods. Positions 421 and 422 are part of the PP domain. The alignment method using the standard numbering scheme was able to provide a perfect match in the structural superimposition of the reference structure (pyruvate decarboxylase from *S. cerevisiae*; pdb|2VK8; gray) and the structures of cyclohexane-1,2-dione hydrolase from *Azoarcus sp.* (AsCDH, pdb|2PGO; orange, a) and the pyruvate decarboxylase from *A. pasteurianus* (ApPDC, pdb|2VBI; green, b). The residues at standard positions 421 and 422 of the ScPDC are colored in red, the corresponding positions in ApPDC and AsCDH are colored in blue.

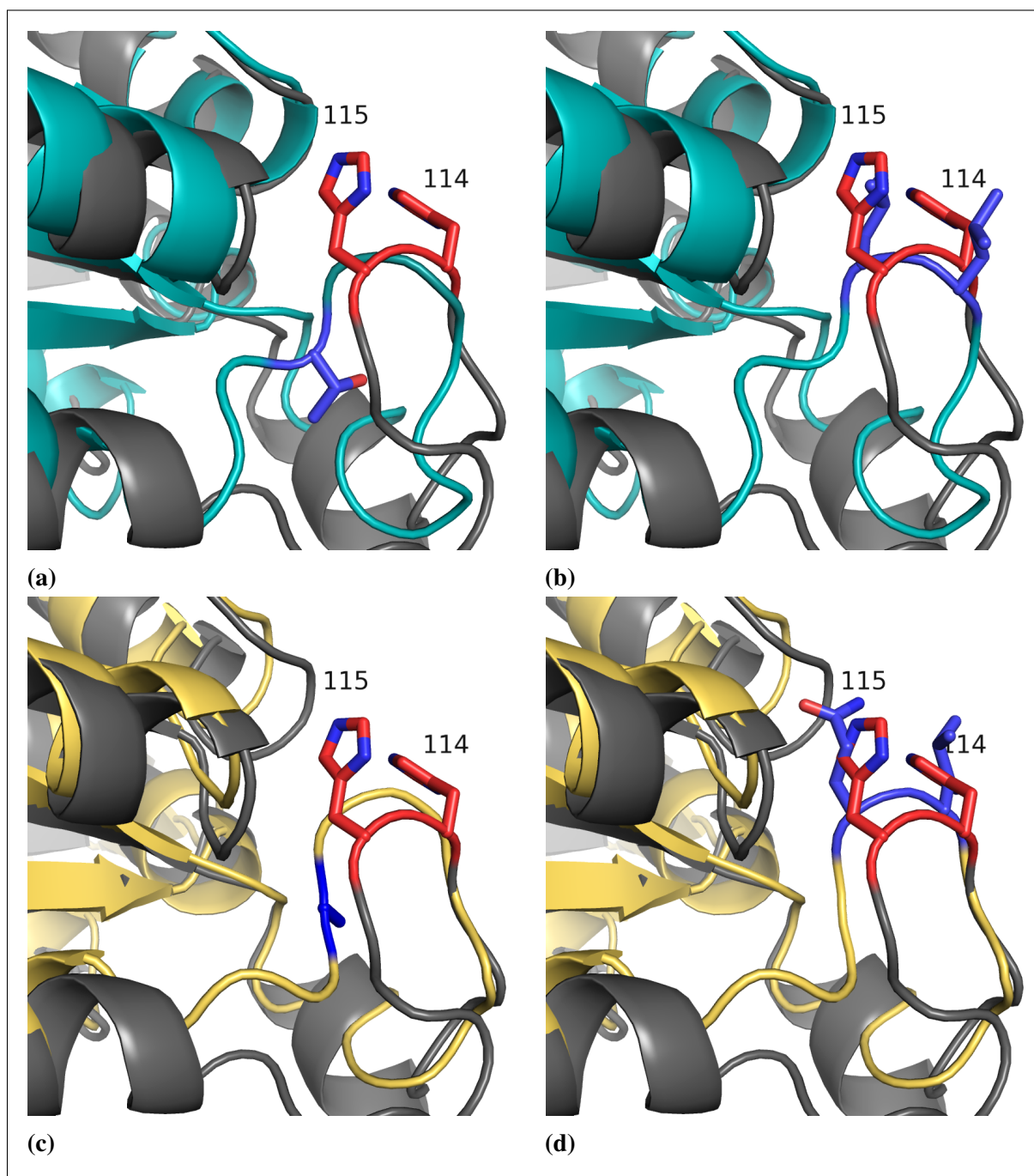


Figure A.3: 'Dissimilar' positions 114 and 115 (HH-motif) of the alignment comparison between T-Coffee and the alignment method using the standard numbering scheme. The numbering scheme based alignment led to an accurate prediction of the residues belonging to the HH-motif of the pyruvate decarboxylases (Figures b and d), while the positions, which were aligned against the positions 114 and 115 of the pyruvate decarboxylase from *Saccharomyces cerevisiae* (*ScPDC*) by T-Coffee, do not show structural correspondence (a and c). The superimposition shows overlays of the reference structure (*ScPDC*; pdb|2VK8; gray) with the benzoylformate decarboxylase from *P. putida* (pdb|1MCZ; cyan, a and b) and with the benzaldehyde lyase from *P. fluorescens* (pdb|3D7K; yellow, c and d). The residues at standard positions 114 and 115 of the *ScPDC* are colored in red, the respective positions predicted by either the standard numbering scheme of T-Coffee are colored in blue.

Table A.3: 22 positions of interest in THDP-dependent decarboxylases including positions with known functional relevance and the assigned domain boundaries. Concerning the function of specific residues also see table A.4 on pages 191ff.

Reference position	6	25	26	27	28	51	114	115	168	197	221	
Function	PYR start	S-pocket	S-pocket	S-pocket	S-pocket	cofactor activation	active site	active site	PYR end	TH3 start	activator binding	
Protein	Acc nr											
<i>AppDC</i>	tr Q8L388	V5	V24	G25	G26	D27	E50	H113	H114	A167	L195	K219
<i>PpBFDc</i>	gi 3915757	V4	N23	P24	G25	S26	E47	L109	L110	P159	N186	D210
<i>EcAHAS I</i>	gi 16131541	G15	I34	P35	G36	G37	E60	F122	Q123	P171	S197	G221
<i>EcAHAS II</i>	gi 33112641	G3	Y22	P23	G24	G25	E47	F109	Q110	P158	P182	G206
<i>PpBAL</i>	gi 1705519	G6	L25	H26	G27	A28	E50	L112	Q113	P162	D190	E214
<i>EcGXC</i>	gi 84028422	A6	V25	P26	G27	A28	V51	F114	Q115	P163	S189	G213
<i>LkdcA</i>	gi 75369656	V4	V23	P24	G25	D26	E49	H112	H113	P165	E191	E215
<i>EcMend</i>	spl P17109	W10	A29	P30	G31	S32	E55	N117	Q118	P169	E207	
Most frequent aa	G 44%	Y 33%	P 74%	G 91%	G 39%	E 94%	F 58%	Q 81%	P 87%	D 18%	G 71%	
	V 13%	I 19%	V 6%	D 18%	V 3%	H 15%	H 12%	A 3%	E 9%	L 4%		
Reference position	336	367	443	444	445	446	471	473	476	477	540	
Function	TH3 end	PP start	GDGX	GDGX	GDGX	GDGX	Mg ²⁺ binding	Mg ²⁺ binding	S-pocket	S-pocket	PP end	
Protein	Acc nr											
<i>AppDC</i>	tr Q8L388	K334	N363	G434	D435	G436	S437	N462	G464	I467	E468	I530
<i>PpBFDc</i>	gi 3915757	L328	P355	G427	D428	G429	S430	N455	T457	A460	L461	T523
<i>EcAHAS I</i>	gi 16131541	L337	H370	G443	D444	G445	S446	N471	A473	L476	V477	I540
<i>EcAHAS II</i>	gi 33112641	P324	A354	G427	D428	G429	S430	N455	R457	M460	V461	I524
<i>PpBAL</i>	gi 1705519	A332	P372	G447	D448	G449	S450	N475	S477	A480	T481	V544
<i>EcGXC</i>	gi 84028422	V332	P372	G445	D446	F447	D448	N473	Y475	L478	I479	L555
<i>LkdcA</i>	gi 75369656	L330	Q356	G428	D429	G430	S431	N456	G458	V461	E462	L526
<i>EcMend</i>	spl P17109	P338	E369	G441	D442	L443	S444	N469	G471	I474	F475	V537
Most frequent aa	L 17%	P 54%	G 98%	D 91%	G 70%	S 52%	N 89%	G 34%	M 42%	V 45%	V 37%	
	A 13%	E 8%	D 7%	A 9%	G 20%	D 9%	Y 12%	I 18%	I 20%	I 22%		

Table A.4: Functionally relevant positions in selected ThDP-dependent decarboxylases found in the literature.

Protein	Organism	PDB	Mutation	Std. pos.	Effect	Reference
PDC	<i>S. cerevisiae</i>	1QPB	D28A,N	28	Is involved in rate limiting steps of decarboxylation; D28N,A catalyzes formation of <i>S</i> -acetolactate as the major product, besides acetoin.	Liu et al. 2001; Sergienko and Jordan 2001b; Sergienko and Jordan 2001a
			E51D,Q,A	51	Stabilizes the cofactor ThDP; E51Q decreases the catalytic activity, E51A leads to inactivity.	Killenber-Jabs et al. 1997; Candy, Koga, et al. 1996
			E91D,Q,A	91	Stabilizes the zwitter-ionic enamine intermediate. 10 ⁹ fold acceleration of decarboxylation of hydroxyl-benzyl-ThDP; part of the substrate activation cascade; charge necessary for activity.	Jordan, Li, and Brown 1999; Li, Furey, and Jordan 1999
			H92	92	Part of the substrate activation cascade.	Li, Furey, and Jordan 1999
			H114F	114	Is involved in rate limiting steps of decarboxylation.	Liu et al. 2001; Sergienko and Jordan 2001b; Sergienko and Jordan 2001a
			H115F	115	Is involved in rate limiting steps of decarboxylation.	Liu et al. 2001; Sergienko and Jordan 2001b; Sergienko and Jordan 2001a
			C221S,A	221	Important residue for allosteric activation by pyruvate or pyruvate amide.	Baburina, Dikdan, et al. 1998; Baburina, Gao, et al. 1994; Joseph et al. 2006; Lu et al. 2000; Wang et al. 2001
			W412F,A	412	Impact on substrate activation, Alavariant: very much reduced substrate activation.	Li, Furey, and Jordan 1999; Li and Jordan 1999

continued Table A.4: Functionally relevant positions in selected ThDP-dependent decarboxylases found in the literature.

Protein	Organism	PDB	Mutation	Std. pos.	Effect	Reference
			E477Q	477	Is involved in rate limiting steps of decarboxylation.	Liu et al. 2001; Sergienko and Jordan 2001b; Sergienko and Jordan 2001a
	<i>A. pasteurianus</i>	2VBI	W388A,I	392	Alanin-variant shows reduced <i>R</i> resp. (<i>S</i>)-selectivity, although <i>S</i> -pocket is closed by E469. Suggests alternative <i>S</i> -pathway.	Rother et al. 2011
			E469G	477	Mutation opens <i>S</i> -pocket and (<i>S</i>)-2-hydroxyketones are formed; almost no decarboxylase activity.	Rother et al. 2011
	<i>Z. mobilis</i>	1ZPD	D27E,N,A	28	Strongly reduced decarboxylase activity; D27A shows weak acetolactate forming activity.	Chang, Nixon, and Duggleby 1999; Huang et al. 2001; Wu et al. 2000
			E50Q	51	Polarizes pyrimidine ring.	Huang et al. 2001
			H113Q,K,R	114	H113Q inactive for decarboxylation.	Huang et al. 2001; Schenk, Leeper, et al. 1997
			H114Q,A	115	H114Q k_{cat} slower.	Huang et al. 2001; Schenk, Leeper, et al. 1997
			W392A,I,M	392	Improved carboligase activity.	Bruhn et al. 1995; Goetz et al. 2001; Iwan et al. 2001; Pohl 1997
			I472A	476	Altered substrate range.	Siegert et al. 2005
			E473D,Q	477	Decarboxylation of α -lactyl-ThDP and protonation of HE-ThDP is reduced, but carboligase activity is increased; cofactor binding in E473Q is tighter than in wtZmPDC	Chang, Nixon, and Duggleby 1999; Huang et al. 2001; Breslow 1957; Meyer, Neumann, Parthier, et al. 2010; Meyer, Walter, et al. 2011
			I476F	480	Altered substrate range.	Siegert et al. 2005

continued Table A.4: Functionally relevant positions in selected ThDP-dependent decarboxylases found in the literature.

Protein	Organism	PDB	Mutation	Std. pos.	Effect	Reference
BAL	<i>P. fluorescens</i>	2AG0	A28S	28	Introduces weak decarboxylase activity into BAL; reduces ligase activity; A28 corresponds to S26 in BFD.	Brandt, Kneen, et al. 2010; Brandt, Nemeria, et al. 2008; Janzen et al. 2006; Kneen, Pogozheva, et al. 2005
BFD	<i>P. putida</i>	1BFD	S26	28	Reduces decarboxylase activity. Increased K_M for benzoylformate. S26 corresponds to A28 in BAL.	Kneen, Pogozheva, et al. 2005
			H281A	292	Improved benzoin forming activity.	Dünkelmann et al. 2002; Kokova et al. 2009
			A460I	476	Altered substrate range.	Siegert et al. 2005
			L461A/G	477	Residue determines size of S-pocket; Mutation decreases decarboxylase activity	Gocke, Walter, et al. 2008

A.4.1 Description of the 'nvw' file format for biological sequences using a reference sequence based standard numbering scheme

In order to provide a file format, which is able to display amino acid sequences and position specific standard numbers, the 'nvw' file format was developed. It contains a header, a title, optional annotation information and the numbered sequence. The header is marked by the '#'-sign and can span several lines. Each 'nvw' file can contain one title, which has to be placed between two '//'-signs. The optional lines starting with '#ANNODESC' and '#ANNOPOS' provide additional information about the position and the description of annotation in the numbered sequence. The '#ANNODESC' lines define the description of annotations and must consist of the following pattern:

```
#ANNODESC [description id] [description] [color]
```

Each annotation description must have a unique description id and a distinct color in hexadecimal color code to allow highlighting of the respective positions in multisequence alignments, which can be generated based on 'nvw'-files. The '#ANNOPOS' lines define the position of given annotations on the numbered sequence and consist of:

```
#ANNOPOS [standard position number] [description id]
```

The sequence and the standard numbers have to be given in vertical columns of maximum 50 rows. The example nvw file 'example_2VK8.nvw' shows the sequence of the *ScPDC* (pdb|2VK8) numbered using the presented numbering scheme (Section A.4.2).

A.4.2 Example nvw-file containing the numbered and annotated sequence of the pyruvate decarboxylase from *S. cerevisiae*

Listing A.1: 'nvw' file of the pyruvate decarboxylase from *S. cerevisiae*. The sequence of the pyruvate decarboxylase from *S. cerevisiae*, which was used as the reference sequence of the numbering scheme for the decarboxylase superfamily of the ThDP-dependent enzymes, was numbered by applying the numbering scheme. The nvw-file 'example_2VK8.nvw' contains the sequence of the respective protein with the standard numbers for each position and 27 position specific annotations.

```
# nvw-numbering
# 2014 ITB Uni Stuttgart Germany

//example_2VK8//

#ANNODESC 0 GDGX-motiv , TPP-binding #00FF33
#ANNODESC 1 PP-end #053BF9
```

```
#ANNODESC 2 PYR-end #FC0C02
#ANNODESC 3 S-pocket #ffdead
#ANNODESC 4 TH3-end #FF8C00
#ANNODESC 5 Mg2+ binding #CC6600
#ANNODESC 6 GDGX-motiv , Mgs+ binding #CC6600
#ANNODESC 7 entrance to S-pocket #ffdead
#ANNODESC 8 activator binding #ffdead
#ANNODESC 9 S-pocket #ffdead
#ANNODESC 10 TH3-start #FF8C00
#ANNODESC 11 GDGX-motiv #FFCCFF
#ANNODESC 12 PYR-start #FC0C02
#ANNODESC 13 ThDP-binding #9932cc
#ANNODESC 14 donor-pocket #ffdead
#ANNODESC 15 PX motif #uu6785
#ANNODESC 16 PP-start #053BF9
```

```
#ANNOPOS 6 12
#ANNOPOS 25 3
#ANNOPOS 26 3
#ANNOPOS 27 3
#ANNOPOS 28 3
#ANNOPOS 51 13
#ANNOPOS 114 14
#ANNOPOS 115 14
#ANNOPOS 166 15
#ANNOPOS 167 15
#ANNOPOS 168 2
#ANNOPOS 169 15
#ANNOPOS 170 15
#ANNOPOS 171 15
#ANNOPOS 197 10
#ANNOPOS 221 8
#ANNOPOS 336 4
#ANNOPOS 367 16
#ANNOPOS 443 11
#ANNOPOS 444 6
#ANNOPOS 445 11
#ANNOPOS 446 0
#ANNOPOS 471 5
#ANNOPOS 473 5
#ANNOPOS 476 7
#ANNOPOS 477 9
#ANNOPOS 540 1
```

```
1 M 51 E 101 V 151 R 201 I 251 S 301 Y 351 A 401 P 451 V 501 T 551 E
2 S 52 L 102 P 152 C 202 D 252 I 302 K 352 R 402 N 452 Q 502 F 552 Q
3 E 53 N 103 S 153 I 203 T 253 D 303 T 353 T 403 N 453 E 503 G 553 A
4 I 54 A 104 I 154 R 204 I 254 E 304 K 354 P 404 T 454 I 504 A 554 K
5 T 55 A 105 S 155 T 205 L 255 Q 305 N 355 A 405 Y 455 S 505 K 555 L
6 L 56 Y 106 A 156 T 206 V 256 H 306 I 356 N 406 G 456 T 506 D 556 T
7 G 57 A 107 Q 157 Y 207 L 257 P 307 V 357 A 407 I 457 M 507 Y 557 A
8 K 58 A 108 A 158 V 208 D 258 R 308 E 358 A 408 S 458 I 508 E 558 A
9 Y 59 D 109 K 159 T 209 K 259 Y 309 F 359 V 409 Q 459 R 509 T 559 T
10 L 60 G 110 Q 160 Q 210 D 260 G 310 H 360 P 410 V 460 W 510 H 560 N
11 F 61 Y 111 L 161 R 211 A 261 G 311 S 361 A 411 L 461 G 511 R 561 A
12 E 62 A 112 L 162 P 212 K 262 V 312 D 362 S 412 W 462 L 512 V 562 K
13 R 63 R 113 L 163 V 213 N 263 Y 313 H 363 T 413 G 463 K 513 A 563 Q
14 L 64 I 114 H 164 Y 214 P 264 V 314 M 364 P 414 S 464 P 514 T
```

15	K	65	K	115	H	165	L	215	V	265	G	315	K	365	L	415	I	465	Y	515	T
16	Q	66	G	116	T	166	G	216	I	266	T	316	I	366	K	416	G	466	L	516	G
17	V	67	M	117	L	167	L	217	L	267	L	317	R	367	Q	417	F	467	F	517	E
18	N	68	S	118	G	168	P	218	A	268	S	318	N	368	E	418	T	468	V	518	W
19	V	69	C	119	N	169	A	219	D	269	K	319	A	369	W	419	T	469	L	519	D
20	N	70	I	120	G	170	N	220	A	270	P	320	T	370	M	420	G	470	N	520	K
21	T	71	I	121	D	171	L	221	C	271	E	321	F	371	W	421	A	471	N	521	L
22	V	72	T	122	F	172	V	222	C	272	V	322	P	372	N	422	T	472	D	522	T
23	F	73	T	123	T	173	D	223	S	273	K	323	G	373	Q	423	L	473	G	523	Q
24	G	74	F	124	V	174	L	224	R	274	E	324	V	374	L	424	G	474	Y	524	D
25	L	75	G	125	F	175	N	225	H	275	A	325	Q	375	G	425	A	475	T	525	K
26	P	76	V	126	H	176	V	226	D	276	V	326	M	376	N	426	A	476	I	526	S
27	G	77	G	127	R	177	P	227	V	277	E	327	K	377	F	427	F	477	Q	527	F
28	D	78	E	128	M	178	A	228	K	278	S	328	F	378	L	428	A	478	K	528	N
29	F	79	L	129	S	179	K	229	A	279	A	329	V	379	Q	429	A	479	L	529	D
30	N	80	S	130	A	180	L	230	E	280	D	330	L	380	E	430	E	480	I	530	N
31	L	81	A	131	N	181	L	231	T	281	L	331	Q	381	G	431	E	481	H	531	S
32	S	82	L	132	I	182	Q	232	K	282	I	332	K	382	D	432	I	482	G	532	K
33	L	83	N	133	S	183	T	233	K	283	L	333	L	383	V	433	D	483	P	533	I
34	L	84	G	134	E	184	P	234	L	284	S	334	L	384	V	434	P	484	K	534	R
35	D	85	I	135	T	185	I	235	I	285	V	335	T	385	I	435	K	485	A	535	M
36	K	86	A	136	T	186	D	236	D	286	G	336	T	386	A	436	K	486	Q	536	I
37	I	87	G	137	A	187	M	237	L	287	A	337	I	387	E	437	R	487	Y	537	E
38	Y	88	S	138	M	188	S	238	T	288	L	338	A	388	T	438	V	488	N	538	V
39	E	89	Y	139	I	189	L	239	Q	289	L	339	D	389	G	439	I	489	E	539	M
40	V	90	A	140	T	190	K	240	F	290	S	340	A	390	T	440	L	490	I	540	L
41	E	91	E	141	D	191	P	241	P	291	D	341	A	391	S	441	F	491	Q	541	P
42	G	92	H	142	I	192	N	242	A	292	F	342	K	392	A	442	I	492	G	542	V
43	M	93	V	143	A	193	D	243	F	293	N	343	G	393	F	443	G	493	W	543	F
44	R	94	G	144	T	194	A	244	V	294	T	344	Y	394	G	444	D	494	D	544	D
45	W	95	V	145	A	195	E	245	T	295	G	345	K	395	I	445	G	495	H	545	A
46	A	96	L	146	P	196	S	246	P	296	S	346	P	396	N	446	S	496	L	546	P
47	G	97	H	147	A	197	E	247	M	297	F	347	V	397	Q	447	L	497	S	547	Q
48	N	98	V	148	E	198	K	248	G	298	S	348	A	398	T	448	Q	498	L	548	N
49	A	99	V	149	I	199	E	249	K	299	Y	349	V	399	T	449	L	499	L	549	L
50	N	100	G	150	D	200	V	250	G	300	S	350	P	400	F	450	T	500	P	550	V

T-Coffee 9.01

1PVD	LLHH	TLGNGDFTV	FHRMSANIS	ETTAMI	TDI	----	ATAPA	----	EIDRCIRTTYV	I	QRPVYLGLPAN	LV	D	----	L	NV	P	----	AK	I	IQ	TP	ID	
3D7K	LQA	----	GIDQVAMA	APITKWAHRV	MAT	----	EHIPR	----	LVMQAIRAALS	A	RGPVLLDLPWD	IL	M	----	N	QI	D	----	ED	S	VI	IP	----	
2NXW	LLHH	QGRTLD	----	TQFQVFKEITVAQARL	DDP	----	AKAPA	----	EIARVLGAARA	Q	SRPVYLEIPRN	MV	N	----	A	EV	E	----	PV	G	DD	----	----	
1N0H	F-Q	----	EADVVGISRS	CTKWNVMV	KSV	----	EELPL	----	RINEAFEIATS	GR	PGPVLVDLPKD	VT	A	----	A	IL	RN	----	PI	PTK	IT	LP	----	
1POX	FQE	----	M	----	NENPIYADVADYNVTA	VNA	----	ATLPH	----	VIDEAIRRAYA	H	QGVAVVQIPVD	LP	W	----	Q	QI	S	----	AE	D	WY	AS	AN
2PGO	AAQ	----	QV	----	PWQSFTPIARSTQRV	ERL	----	DKVGE	----	AIHEAFRVAEGH	P	AGPAYVDIPFD	LT	A	----	D	QI	D	----	DK	A	LV	PR	----
3FLM	NQ	----	----	ATRQPGMFASHPTHSISL	PRPT	----	QDIPARWLVS	----	TIDHALG	----	TLH	AGGVHINC	PFA	EPLY	----	G	EM	D	----	DT	G	LS	WQ	----
3EA4	F-Q	----	ET	PIVEVTRSITKHNYLV	MDV	----	EDIPR	----	IIEEAFFLATS	GR	PGPVLVDVPKD	IQ	Q	----	Q	LA	I	----	PN	WEQ	AM	RI	----	
1UPA	H-Q	----	CL	DSVAIVAPMSKYAVEL	QRP	----	HEITD	----	LVDSAVNAAMTE	P	VGPSFISLPVD	LL	G	----	S	SE	G	----	ID	TI	VP	NPAN	----	
1MCZ	LLT	----	NV	DAANLPRPLVKWSYEP	ASA	----	AEVPH	----	AMSRAIHMAS	MAP	QGPVYLSVPYD	DW	D	----	K	DA	D	----	PQ	S	HH	LF	----	
3EY9	F-Q	----	ET	HPQELFREC	SHYCELV	SSP	----	EIQIPQ	----	VLAIAMRKAVL	N	RGVSVVVLPGD	VA	L	----	K	PA	P	----	EG	A	TM	H	----
2Q28	Y-E	----	EL	QMNAAKPYAKAAFRV	NQP	----	QDLGI	----	ALARAIRVSVS	GR	PGGVYLDLPAN	VL	A	----	A	TM	E	----	KD	EA	TI	LV	----	
1OZG	H-Q	----	SM	DTVAMFSPVTKYAIEV	TAP	----	DALAE	----	VVSNAFRAAEQ	GR	PGSAFVSLPQD	VV	D	----	G	PV	S	----	GK	V	LP	----	----	
2VBG	FVHH	TLADGDFK	HFMKMHEPV	TAARTLL	TA	----	ENATY	----	EIDRVLSQLK	E	RKPVYINLPVD	VA	A	----	A	KA	E	----	KP	A	LS	LE	----	
2VBI	LLHH	TI	GKTDYSYQLE	MARQVTCAAESI	TDA	----	HSAPA	----	KIDHVIRTALR	E	RKPAYLDIACN	IA	S	----	E	PC	V	----	RP	G	PV	SS	----	

Standard numbering scheme

1PVD	LLHH	TLGNGDFTV	FHRMSANIS	ETTAMI	TDI	----	ATAPA	----	EIDRCIRTTYV	I	QRPVYLGLPAN	LV	D	----	L	NV	P	----	AK	I	IQ	TP	ID		
3D7K	NTLQAGI	----	----	DQVAMAAPITKWAHRV	MAT	----	EHIPR	----	LVMQAIRAALS	A	RGPVLLDLPWD	IL	M	----	N	QI	DEDS	----	VI	I	P	DI	VI		
2NXW	LLHH	QGRT	----	LDTQFQVFKEITVAQARL	DDP	----	AKAPA	----	EIARVLGAARA	Q	SRPVYLEIPRN	MV	N	----	A	EV	E	----	P	V	CD	DP	AW		
1N0H	DAFQ	EA	----	----	DVVGISRSCTKWNVMV	KSV	----	EELPL	----	RINEAFEIATS	GR	PGPVLVDLPKD	VT	A	AILRN	PI	PTKT	----	TI	P	SNAL	NQ	LT		
1POX	DTFQ	EM	----	----	NENPIYADVADYNVTA	VNA	----	ATLPH	----	VIDEAIRRAYA	H	QGVAVVQIPVD	LP	W	----	Q	QI	S	A	ED	W	Y	AS	AN	
2PGO	EAAQ	QV	----	----	PWQSFTPIARSTQRV	ERL	----	DKVGE	----	AIHEAFRVAEGH	P	AGPAYVDIPFD	LT	A	----	D	QI	DD	K	AL	V	P	R	GA	TR
3FLM	GANC	AI	----	----	RQPGMFASHPTHSISL	PRPT	QDIPARWLVS	----	TIDHALG	TL	H	AGGVHINC	PFAEPL	Y	----	G	EM	DDTGLSWQRL	GD	W	WQDD	KP	WL		
3EA4	DAFQ	ET	----	----	PIVEVTRSITKHNYLV	MDV	----	EDIPR	----	IIEEAFFLATS	GR	PGPVLVDVPKD	IQ	Q	----	Q	LA	I	----	PN	W	EQAMR	LP	GY	
1UPA	DIHQ	CL	----	----	DSVAIVAPMSKYAVEL	QRP	----	HEITD	----	LVDSAVNAAMTE	P	VGPSFISLPVD	LL	G	----	S	SEGID	----	TI	VP	NP	PA	NT		
1MCZ	EALL	IN	----	----	VDAANLPRPLVKWSYEP	ASA	----	AEVPH	----	AMSRAIHMAS	MAP	QGPVYLSVPYD	DW	D	----	K	DA	D	P	QS	H	HLF	DR	HV	
3EY9	GYFQ	ET	----	----	HPQELFREC	SHYCELV	SSP	----	EIQIPQ	----	VLAIAMRKAVL	N	RGVSVVVLPGD	VA	L	----	K	PA	PE	G	AT	M	H	WY	HA
2Q28	GDIY	EL	----	----	QMNAAKPYAKAAFRV	NQP	----	QDLGI	----	ALARAIRVSVS	GR	PGGVYLDLPAN	VL	A	----	A	TM	EKDEAL	TT	I	V	KV	EN		
1OZG	QVHQ	SM	----	----	DTVAMFSPVTKYAIEV	TAP	----	DALAE	----	VVSNAFRAAEQ	GR	PGSAFVSLPQD	VV	D	----	G	PV	S	----	GK	V	LP	AS	GA	
2VBG	FVHH	TLADGDFK	HFMKMHEPV	TAARTLL	TA	----	ENATY	----	EIDRVLSQLK	E	RKPVYINLPVD	VA	A	----	A	KA	E	----	KP	A	LS	L	EK	ES	
2VBI	LLHH	TI	GKTDYSYQLE	MARQVTCAAESI	TDA	----	HSAPA	----	KIDHVIRTALR	E	RKPAYLDIACN	IA	S	----	E	PC	VR	P	----	GP	V	S	SL	LSE	

reference sequence

2VK8	LLHH	TLGNGDFTV	FHRMSANIS	ETTAMI	TDI	----	ATAPA	----	EIDRCIRTTYV	I	QRPVYLGLPAN	LV	D	----	L	NV	P	----	AK	I	IQ	TP	ID		
sim.	2222	022	5544	22	444	5555	55555555	55555555	55555555	444	000000	44444	00000	55555555	445	0000	4044	4000000000	22	00	20	23	000	22	0110

T-Coffee 9.01

1PVD	M	S	LK	PN	D	A	ESEK	EVIDTILALVKDAKNPVILADACCSRHD	VKAETKKLIDLTQFPAFVTPMG	KGSI	SEQ	HPRYG	GVVGT	LSK		
3D7K	D	L	VL	SAH	G	A	RPDP	ADLDQALALLRKAERPVIIVLGSEASRTA	RKTALSAFVAATGVPVFADYEG	LSML	SGL	PDAMR	GGL	VQNL		
2NXW	I	P	AW	PV	D	R	DALA	ACADEVLAAMRSATSPVLMVCVEVRRYG	LEAKVAELAQRLLGVPVVTTFMG	RGLL	ADA	PTPPL	GTYIGV	AGD		
1N0H	S	N	AL	NQL	T	S	RAQDEFVM	QSINKAADLINLAKKPVLYVGAGILNHAD	GPRLKELSDRAQIPVTTTLQG	LGSF	DQE	DPKSL	DM	LGM	HGC	
1POX	N	Y	QT	PL	L	P	EPDV	QAVTRLTQTLAAERPLIYYGIGARKA	GKELEQLSKTLKIPLMSTYPA	KGIV	ADR	YPAYL	GSA	NR	VAQ	
2PGO	G	A	TR	AKS	V	L	HAPN	EDVREAAAQLVAAKNPVILAGGGVARSG	GSEALLKLAEMVGVVVTSTG	AGVF	PET	HALAM	GSA	GF	CGW	
3FLM	QRLGDW	W	QD	DKP	W	LRE	APRL	ESEKQRDWFWRQKRGVVVAGRMSAEEG	K--KVALWAQTLGWPLIGDVL	Q---	TGQ	PLPCA	DLW	LGN		
3EA4	P	G	YM	SRM	P	K	PPED	SHLEQIVRLISESKKPVLYVGGGCLNSS	DELGRFVELTGIPVATLMG	LGSY	PCD	DELSL	HM	LGM	HGT	
1UPA	I	P	AK	PV	GVV	A	DGWQ	KAADQAAALLAEAKHPVLVVGAAAIRSG	AVPAIRALAERLNIPVITTYIA	KGVL	PVG	HELNY	GAVT	GYMDGI	LNF	
1MCZ	D	R	HV	SS	V	V	RLND	QDLDILVKALNSASNPAIVLGPDVDAAN	ANADCVMLAERLKAPVWVAPSAP	RCPF	PTR	HPCFR	GL	M	P	AGI
3EY9	W	Y	HA	PQ	V	V	TPEE	EELRKLQALLRYSSNIALMCGSGCAG	AHKELVEFAGKIKAPIVHALRG	KEHV	EYD	NPYDV	GM	TGL	IGF	
2Q28	K	V	EN	PS	P	A	LPCP	KSVTSAISLLAKAERPLIILGKGAAYSQ	ADEQLREFIESAQIPFLPMSMA	KGIL	EDT	HPLSA	A	---	---	
1OZG	A	S	GA	PQ	M	G	AAPD	DAIDQVAKLIAQAKNPIFLLGLMASQPE	NSKALRRLLETSHIPVTSTYQA	AGAV	NQD	NFSRFA	GR	VGL	FNN	
2VBG	I	K	ES	SI	T	N	ITEQ	VILSKIEESLKNAPVVIAGHEVISFG	LEKTVTQFVSETKLPITTLNFG	KSAV	DES	LPSFL	GI	YNGK	LSE	
2VBI	L	L	SE	PEI	D	H	ISLK	AAVDATVALLEKSASPVMLLGSKLRAAN	ALAAATETLADKLQCAVTIMAAA	KGFF	PED	HAGFR	GL	YWGE	VSN	

Standard numbering scheme

1PVD	M	S	LK	PN	D	A	ESEK	EVIDTILALVKDAKNPVILADACCSRHD	VKAETKKLIDLTQFPAFVTPMG	KGSI	SEQ	HPRYG	GVVGT	LSK			
3D7K	S	A	HGAR				PDP	ADLDQALALLRKAERPVIIVLGSEASRTA	RKTALSAFVAATGVPVFADYEG	LSML	SGL	DAMRG	GL	VQNL	YSF		
2NXW	P	V	D	RD			ALA	ACADEVLAAMRSATSPVLMVCVEVRRYG	LEAKVAELAQRLLGVPVVTTFMG	RGLL	ADA	PTPPL	GTYIGV	AGD			
1N0H	S	R	A	QD			EFVM	QSINKAADLINLAKKPVLYVGAGILNHAD	GPRLKELSDRAQIPVTTTLQG	LGSF	DQE	DPKSL	DM	LGM	HGC		
1POX	N	Y	QTP	LI	PE		PDV	QAVTRLTQTLAAERPLIYYGIGARK	AGKELEQLSKTLKIPLMSTYPA	KGIV	ADR	YPAYL	GSA	N	VAQ		
2PGO	A	K	SVLH				APN	EDVREAAAQLVAAKNPVILAGGGVARSG	GSEALLKLAEMVGVVVTSTG	AGVF	PET	HALAM	GSAG	F	CGW		
3FLM	R	E	A	PR			LES	EKQRDWFWRQ--KRGVVVAGR-MS-AE	EGKKVALWAQTLGWPLIGDVL	---	SQT	GQP	LPC	A	D	LW	LGN
3EA4	M	S	RMP	KP			PED	SHLEQIVRLISESKKPVLYVGGGCL--N	SSDELGRFVELTGIPVATLMG	LGSY	PCD	DELSL	HM	LGM	HGT		
1UPA	P	A	K	PV	GVVA		DGWQ	KAADQAAALLAEAKHPVLVVGAAAIRSG	AVPAIRALAERLNIPVITTYIA	KGVL	PVG	HELNY	GAVT	GYMDGI	LNF		
1MCZ	S	S	S	VR			LND	QDLDILVKALNSASNPAIVLGPDVDAAN	ANADCVMLAERLKAPVWVAPSAP	RCPF	PTR	HPCFR	GL	M	P	AGI	
3EY9	P	Q	PV	VI			PEE	EELRKLQALLRYSSNIALMCGSGCA--G	AHKELVEFAGKIKAPIVHALRG	KEHV	EYD	NPYDV	GM	TGL	IGF		
2Q28	P	S	P	A	L		LPCP	KSVTSAISLLAKAERPLIILGKGAAYSQ	ADEQLREFIESAQIPFLPMSMA	KGIL	EDT	HPLSA	AAA	---	---		
1OZG	P	Q	M	GA			APD	DAIDQVAKLIAQAKNPIFLLGLMASQPE	NSKALRRLLETSHIPVTSTYQA	AGAV	NQD	NFSRFA	GR	VGL	FNN		
2VBG	S	I	T	NT			TEQ	VILSKIEESLKNAPVVIAGHEVISFG	LEKTVTQFVSETKLPITTLNFG	KSAV	DES	LPSFL	GI	YNGK	LSE		
2VBI	P	E	ID	HT			SLK	AAVDATVALLEKSASPVMLLGSKLRAAN	ALAAATETLADKLQCAVTIMAAA	KGFF	PED	HAGFR	GL	YWGE	VSN		

reference sequence

2VK8	M	S	LK	PN	D	A	ESEK	EVIDTILVLDKDAKNPVILADACCSRHD	VKAETKKLIDLTQFPAFVTPMG	KGSI	DEQ	HPRYG	GVVGT	LSK																													
sim.	Z	00000	I	000	12	22	00	I	00	I	00	2	444	0000	5	4444444444444444	5	5555555555555555	4	444	3	4	0	444	5	5555555555555555	5	5555	4	4444	0	4444	0	0000	5	3	22	34	0	0000	4	444	0

T-Coffee 9.01

1PVD	PEVKEAVESADLILSVGALL	SDFNT	GSFSY	SY	KT	-----	K	-----	NIVEFSDH	-----	MKIRNATFPGVQM	-----	KFVLQKLLTNIAD	-----	AA	-----	K
3D7K	YSFAKADAAPDLVLMGARF	GLNTGH	HSGQL	IP	HS	-----	A	-----	QVIQVDPDA	-----	CELGRLQGIALGI	VADV	GGTIEALAQATAQ	-----	DA	-----	A
2NXW	AEITRLVEESDGLFLLGAIL	SDTNF	AVSQR	KI	DL	-----	R	-----	KTIHAFDRA	-----	VTLGYHTYADIPL	-----	AGLVDALLERLPP	-----	SD	-----	R
1N0H	ATANLAVQNADLIIAVGARF	DDRVT	GNISK	FA	PE	ARRAAA	EGRG	-----	GIIHFEVSP	-----	KNINKVVQIQIAV	EGDA	TTNLGKMMSKIFP	-----	VK	-----	E
1POX	KPANEALAQADVFLVGNYY	PFAEV	SK--A	FK	NT	-----	R	-----	YFLQIDIDP	-----	AKLGKRHKTDIAV	LADA	QKTLAAILAQVSE	-----	RE	-----	S
2PGO	KSANDMMAAADFVLVLGSRL	SDWGI	AQG-Y	IT	KM	-----	P	-----	KFVHVDIDP	-----	AVLGTFFYFPLLSV	VADA	KTFMEQLIEVLPG	-----	TS	-----	G
3FLM	AKATSELQQAQIVVQLGSSL	TGKRL	LQWQA	SC	EP	-----	E	-----	EYWIIDDIE	-----	GRLDPAHHRGRRI	IANI	ADWLELH----	-----	--	-----	P
3EA4	VYANYAVEHSDLLLAFGVRF	DDRVT	GKLEA	FA	SR	-----	A	-----	KIVHIDIDS	-----	AETGKNKTPHVSV	CGDV	KLALQGMNKVLEN	-----	RA	-----	E
1UPA	PALQTMFAPVDLVLTVGYDY	AEDLRP	SMWQ-	KG	IE	-----	K	-----	KTVRISPTV	-----	NPIPRVYRPDQDV	VTDV	LAFVEHFETATAS	-----	FGAKQRH	-----	D
1MCZ	AAISQLLEGHDDVVLVIGAPV	FRYHQY	DPGOY	LK	PG	-----	T	-----	RLISVTCDP	-----	LEAARAP-MGDAT	VADI	GAMASALANLVEE	-----	SS	-----	R
3EY9	SSGFHTMMNADTLVLLGTQF	PYRA-	----F	YP	TD	-----	A	-----	KIIQIDINP	-----	ASIGAHSKVDMAL	VGDI	KSTLRALLPLVEE	-----	KA	-----	D
2Q28	AARSFALANADVVLVGARL	NWLLA	HGKKG	WA	AD	-----	T	-----	QFIQLDIEP	-----	QEIDSNRPIAVPV	VGDI	ASSMQGMLAELKQ	-----	NT	-----	F
1OZG	QAGDRLLQLADLVICIGYSP	VEYE-	PAMWN-	S-	GN	-----	A	-----	TLVHIDVLP	-----	AYEERNYTPDVEL	VGDI	AGTLNKLAQNIIDH	-----	RL	-----	V
2VBG	ISLKNFVESADFILMLGVKL	TDSST	GAFTH	HL	DE	-----	N	-----	KMISLNIDE	-----	GIIFNKVVEDEFD	-----	RAVVSSLSELKGI	-----	EY	-----	E
2VBI	PGVQELVETSDALLCIAPVF	NDYST	VGWSA	WP	KG	-----	P	-----	NVILAEPDR	-----	VTVDGRAYDGFIL	RA--	FLQAL----	AE	KA	-----	P

Standard numbering scheme

1PVD	PEVKEAVESADLILSVGALL	SDFNT	GSFSY	SY	KT	-----	K	-----	NIVEFSDH	-----	MKIRNATFPGVQM	-----	KFVLQKLLTNIAD	-----	AA	-----	K
3D7K	---AKADAAPDLVLMGARF	GL-NT	HGSG	QL	IPHS	-----	A	-----	QVIQVDPDA	-----	CELGRLQGI-ALGIV-ADV	-----	GGTIEALAQATAQD	-----	AA	-----	W
2NXW	AEITRLVEESDGLFLLGAIL	SDTNF	A-VSQR	KI	DL	-----	R	-----	KTIHAFDRA	-----	VTLGYHTYA-DIPLA-GLV	-----	DALLERL-PPS-D	-----	--	-----	R
1N0H	ATANLAVQNADLIIAVGARF	DDRVT	GNISK	FA	PE	ARRAAA	EGRG	-----	GIIHFEVSP	-----	KNINKVVQT-QIAVE-GDA	-----	TTNLGKMMSKI-F	-----	-P	-----	V
1POX	KPANEALAQADVFLVGNYY	P---F	AEVSK	AF	KNT	-----	R	-----	YFLQIDIDP	-----	AKLGKRHKTDIAVL-ADA	-----	QKTLAAILAQVS	-----	--	-----	E
2PGO	KSANDMMAAADFVLVLGSRL	SDWGI	A--QG	YI	TKM	-----	P	-----	KFVHVDIDP	-----	AVLGTFFYFPLLSV-ADA	-----	KTFMEQLIEVLPG	-----	TS	-----	AV
3FLM	AKATSELQQAQIVVQLGSSL	TGKRL	LQWQA	SC	EP	-----	E	-----	EYWIIDDIE	-----	GRLDPAHHR-GRRLI-ANI	-----	ADWLELH--PA-E	-----	--	-----	K
3EA4	VYANYAVEHSDLLLAFGVRF	DDRVT	GKLEA	FA	SR	-----	A	-----	KIVHIDIDS	-----	SAEIGKNKTP-HVSVC-GDV	-----	KLALQGMNKVLEN	-----	RAE	-----	K
1UPA	PALQTMFAPVDLVLTVGYDY	AEDLR	--PSM	WQ	KGIE	-----	K	-----	KTVRISPTV	-----	NPIPRVYRP-DVDVV-IDV	-----	LAFVEHFETATAS	-----	--	-----	F
1MCZ	AAISQLLEGHDDVVLVIGAPV	FRYHQY	D-PGQ	YL	KPG	-----	T	-----	RLISVTCDP	-----	LEAARAPM--GDAIV-ADI	-----	GAMASALANLV-E	-----	--	-----	E
3EY9	SSGFHTMMNADTLVLLGTQF	PYR--	----A	FY	PTD	-----	A	-----	KIIQIDINP	-----	PASIGAHSKV-DMALV-GDI	-----	KSTLRALLPLVEE	-----	--	-----	K
2Q28	--RSFALANADVVLVGARL	NW-LL	AHGKKG	WA	AD	-----	T	-----	QFIQLDIEP	-----	QEIDSNRPI-AVPVV-GDI	-----	ASSMQGMLAELKQ	-----	NT	-----	F
1OZG	QAGDRLLQLADLVICIGYSP	V--EY	E-PAM	WN	SGN	-----	A	-----	TLVHIDVLP	-----	PAYEERNYTP-DVELV-GDI	-----	AGTLNKLAQNI-D	-----	-H	-----	R
2VBG	ISLKNFVESADFILMLGVKL	TDSST	GAFTH	HL	DE	-----	N	-----	KMISLNIDE	-----	GIIFNKVVEDEFD	-----	RAVVSSLSEL-KG	-----	--	-----	I
2VBI	PGVQELVETSDALLCIAPVF	NDYST	VGWSA	WP	KG	-----	P	-----	NVILAEPDR	-----	VTVDGRAYDGFIL	-----	RAFLQALAEKAPA	-----	--	-----	R

reference sequence

2VK8	PEVKEAVESADLILSVGALL	SDFNT	GSFSY	SY	KT	-----	K	-----	NIVEFSDH	-----	MKIRNATFPGVQM	-----	KFVLQKLLTTIAD	-----	AA	-----	K
sim.	44455555555555555555	0443330	333320	3300	55	0000000000	4	0000000000	55555555	000000	12212222222222	0000	4454445444333	00000	22	00000	2

T-Coffee 9.01

1PVD	-----	G Y K	P	V A V P	A R	T P A N A A	----	V P	A S	T P L K Q E W M W N Q L G N F L	----	Q	----	E	----	G	----	D V V I A E T G T S A F G I N	Q T	T	F P N N	T Y		
3D7K	WP----	A Q E	R	Y A S I	--	A A K	----	S S	S E	H A L H P F H A S Q V I A K H V	----	D	----	A	----	G	----	V T V V A D G A L T Y L W L S	E V	M S R V K P	G G			
2NXW	T-----	G K E	P	H A Y P	T	G L	----	Q A	D G	E P I A P M D I A R A V N D R V	----	R	----	A	----	G O E P	----	L L I A A D M G D C L F T A M	D M	I	D	A G		
1N0H	RS-----	W F A Q I	N K	W K K	E	Y P Y A	Y	M E	----	E T	P G	S K I K P Q T V I K K L S K V A	----	N	----	D T G R H	----	V I V T T G V G Q H Q M W A A	Q H	W T W R N P	H T			
1POX	TP-----	W W Q A N L	A N	V K N	W	R A Y L	A	S L E	----	D K	Q E	G P L Q A Y Q V L R A V N K I A	----	E	----	P	----	A I Y S I D V G D I N L N A N	R H	L K L T P S	N R			
2PGO	FKAVRYQERENFR-QA-TEF	R A A	W	D G W V	--	R E Q	----	S G	D G	M P A S M F R A M A E V R K V Q	----	R	----	P	----	E	----	D I I V T D I G N H T L P M F	G G	A I L Q R P	R R			
3FLM	AE----	A E Q	A M	Q A V I	--	A R	----	R	----	D A F G E A Q L A H R I C D Y L	----	P	----	E Q	----	G	----	Q L F V G N S L V	--	R L I	D A	L S Q	L P	A G
3EA4	ELKL---DFGVWRNEL-NV-	Q K Q	K	F P	--	L	S F	----	K T	F G	E A I P P Q Y A I K V L D E L T	----	D	----	G	----	K	----	A I I S T G V G Q H Q M W A A	Q F	Y N Y K K P	R Q		
1UPA	IEP-----LRARI-AE-	F	--	L A	D	----	P	----	E T	Y E	D G M R V H Q V I D S M N T V M	----	E E A A E P	----	E	----	G T I V S D I G F F R H Y G V	L F	A R A D Q P	F G				
1MCZ	QLP-----	T	A	A P E P	A	K V	----	D Q	D A	G R L H P E T V F D T L N D M A	----	P	----	E	----	N	----	A I Y L N E S T S T T A Q M W	Q R	L N M R N P	G S			
3EY9	RK-----FLDKAL-ED-	Y R D	A	R K G L	D	D L A	----	K P	S E	K A I H P Q Y L A Q I S H F A	----	A	----	D	----	D	----	A I F T C D V G T P T V W A A	R Y	L K M N G K	R R			
2Q28	TTPL-----VWRDIL-NI-	H K Q	Q	N A Q K	M	H E K L	----	S T	D T	Q P L N Y F N A L S A V R D V L	----	R	----	E N	----	Q	----	D I Y L V N E G A N T L D N A R N I	N I	D M Y K P	R R			
1OZG	LSP-----QAAEILRD-	R O H	Q	R E L L	D	R R G	----	Q L	N Q	F A L H P L R I V R A M Q D I V	----	N	----	S	----	D	----	V T L T V D M G S F H I W I A	R Y	L T F R A	R Q			
2VBG	-----	G O	--	Y	I D	K Q Y E E F	----	I P	S S	A P L S Q D R L W Q A V E S L T	----	Q	----	S	----	N	----	E T I V A E Q G T S F F G A S	T I	F	L K S N	S R		
2VBI	AR-----PAS-----	A Q K S	S	V P T C	S	L T A	----	T S	D E	A G L T N D E I V R H I N A L L	----	T	----	S	----	N	----	T T L V A E T G D S W F N A M	R M	L P	R G	A R		

Standard numbering scheme

1PVD	-----	G Y K	P	V A V P	A R	T P A N A A	----	V P	A S	T P L K Q E W M W N Q L G N F L	----	Q	----	E	----	G	----	D V V I A E T G T S A F G I N	Q T	T	F P N N	T Y			
3D7K	-----	P D R G	D W C A K V T D L A Q E R Y A S I A A K S	S S	E H	A L H P F H A S Q V I A K H V	----	D	----	A	----	G	----	V T V V A D G A L T Y L W L S	E V	M S R V K P	P G								
2NXW	-----	T	--	R G K E P	--	H A Y P	G L	Q A	D G	E P I A P M D I A R A V N D R V R A G	----	Q	----	E	----	P	----	L L I A A D M G D C L F T A M	D M	I	D	A G			
1N0H	-----	K E R S	E W F A Q I N K W K K E Y P Y A Y M E	E T	P G	S K I K P Q T V I K K L S K V A N D T	----	G	----	R	----	H	----	V I V T T G V G Q H Q M W A A	Q H W	T	W	R N	P H						
1POX	-----	R E S	T P W W Q A N L A N V K N W R A Y L A S L E	D K	Q E	G P L Q A Y Q V L R A V N K I A	----	E	----	P	----	D	----	A I Y S I D V G D I N L N A N	R H L	K	L	T P S	N R						
2PGO	-----	Y O E R	E N F R Q A T E F R A A D G W V R E Q E	S G	D G	M P A S M F R A M A E V R K V Q	----	R	----	P	----	E	----	D I I V T D I G N H T L P M F	G G A T	L	Q R	P R							
3FLM	-----	R O	P W C V E I P R L A E Q A M Q A V I A R R	D	--	A F G E A Q L A H R I C D Y L	----	P	----	E	----	Q	----	Q L F V G N S L V R L I D	A L S	Q	L P A G	Y P							
3EA4	-----	L D F	G V W R N E L N V Q K Q K F P L S F	K T	F	G	E A I P P Q Y A I K V L D E L T	----	D	----	G	----	K	----	A I I S T G V G Q H Q M W A A	Q F Y	N	Y	K K	P R					
1UPA	-----	G A K	Q	R H D I	--	E P L R A R I A E F L A D P	E T Y E D G M R V H Q V I D S M N T V M E E A A E P	----	G	----	E	----	G T I V S D I G F F R H Y G V	L F A	R	A	D Q	P F							
1MCZ	-----	S S R	Q	--	L P	T A A P E P	--	A	K V	D Q	D A	G R L H P E T V F D T L N D M A	----	P	----	E	----	N	----	A I Y L N E S T S T T A Q M W	Q R L	N	M	R N	P G
3EY9	-----	A D R K F L	D K A L E D	Y R D	A R K G L D	D L A	K P	S E	K A I H P Q Y L A Q I S H F A	----	A	----	D	----	D	----	A I F T C D V G T P T V W A A	R Y L	K	M	N G	K R			
2Q28	-----	T T P L	V W R D I L N I H K Q N A Q K M E K L	S T	D T	Q P L N Y F N A L S A V R D V L R	E	N	----	Q	----	D	----	I Y L V N E G A N T L D N A R	N I	D	M	Y K	P R						
1OZG	-----	L V L	S P Q A A E I L R D R O H Q R E L L D	R R G	A Q L N Q	F A L H P L R I V R A M Q D I V	----	N	----	S	----	D	----	V T L T V D M G S F H I W I A	R Y L	Y	T								

A.4.4 Reference alignment for the standard numbering method

The reference alignment for the numbering method was generated from 16 representative protein structures and is available in the 'clustal' alignment format.

CLUSTAL W(1.60) multiple sequence alignment

```

2JLC  --FNR-RWAAVILEALTRHGVRHICAPGSRSTPLTAAAENS-AF-IHHTHFDERGLGHLALGLAKVSK
1ZPD  MS--Y-TVGYLAERLVQIGLKHFFAVAGDYNLVLLDNLNLLNK-NM-EQVYCCNELNCGFSAEGYARAKG
2VBF  M---Y-TVGDYLLDRLHELGIIEIFGVPGDYNLQFLDQIISRE-DM-KWIGNANELNASYMADGYARTKK
10VM  RTP-Y-CVADYLLDRLTDCGADHLFGVPGDYNLQFLDHVIDSP-DI-CWVGCANELNASYAADGYARCKG
2VK8  MSE-I-TLGKYLFERLQQVNVNTVFGFLPGDFNLSLDDKIYEVE-GM-RWAGNANELNAAAYADGYARIKG
2IHT  --K-P-TAAHALLSRLRDHGVGKVFVGGASMEIHALTRSS-SI-RNVLP RHEQGGVFAAEGYARSSG
1BFD  M-A---SVHGTTYELLRRQGITVFGNPGSNELPFLKDF--PE-DF-RYILALQEACVVGADGYAQAQR
2PGN  MAI-K-RGADLIVEALEEYQTEQVVGFIHTSHFVADAFSKSH-LGKRVINPATELGGAWMVNGYNYVKD
1JSC  V-G-L-TGGQIFNEMMSRQNVDTVFGYPGGAALPVYDAIHNSD-KF-NFVLPKHEQGAGHMAEGYARASG
1YBH  DQP-R-KGADILVEALERQGVETVFAYPGGASMEIHALTRSS-SI-RNVLP RHEQGGVFAAEGYARSSG
2PAN  MAK---RAVDAA-YVLEKEGITTAFGVPGAAPNFYSA-RKHG-GI-RHILARHVEGASH-AEGYTRATA
10ZF  VRQ-WAHGADLVVSQLAQQVRFVFGIPGAKIDKVFDSLDDSS--I-RIIPVRHEANAAMAAVGRITG
2C31  VEL-T-DGFHVLIDALKMNDIDTMYGVVGIPTNLARMWQDDG--Q-RFYSFRHEQHAGYAAIAGYIEG
2AG0  MAM-I-TGGELVVRTLIKAGVEHLFGLGHAIHDTIFQAQLDHD--V-PIIDTRHEAAAAGHAAEGYARAGA
1V5F  -NK-I-NIGLAVMKILESWGADTIYGIPTSLSSLM DAMGEEENNV-KFLQVKHEEVGAMAAVMQSKFPGG
3EYA  -MK-Q-TVAAYIAKTLESAGVKRIWGTGDSLNLGSLNRMG-TI-EWMSTRHEEVAFAAGAEALQSG

2JLC  -QP-VAVIVTSGTAVANLYPALIEAGLTGKELILLTADRPPPEL--I--DC--GANQ-AI-----RQPGM
1ZPD  --A-AAAVVTYSVGLSAFDAIGGAYAENLPVILISGAPNNND--HAAGH--VLHH-ALGKTDYHYQLEM
2VBF  --A-AAFLTTFVGVGELSAINGLAGSYAENLPVVEIVGSPTSKV--QNDGK--FVHH-TLADGDFKHFMMK
10VM  --F-AALLTTFVGVGELSAMNGIAGSYAEHVPVLHIVGAPGTAA--QORGE--LLHH-TLGDGEFRHFYHM
2VK8  --M-SCIITTFVGVGELSALNGIAGSYAEHVGLHVGVPSISA--QAKQL--LLHH-TLNGDFTVFFHRM
2IHT  -RP-QACWATLPGP-TNLSTGIATSVLDRSPVIALAAQSESHD--I---FPNDTHQ-CL-----DSVAI
1BFD  -KP-AFINLHSAAGTGNAMGALSNAWNHSPLIVTAGQQTRAM--I--GV--EALL-TN-----VDAANL
2PGN  -RSAAVGAWHCV-GNLLLHAAMQEARTGRIPAVHIGLNSDGRL--A--GR-SEAAQ-QV-----PW-QS
1JSC  -KP-GVVLVTSVGPATNVVTPMADAFADGIPMVVFTGQVPTSA--I--GT--DAFQ-EA-----DVGVI
1YBH  -KP-GICIAATSGPGATNLVSGLADALLDSVPLVAITGQVPRRM--I--GT--DAFQ-ET-----PIVEV
2PAN  GNI-GVCLGTSGPAGTD-ITALYASADSIPILCITGQAPRAR--L--HK--EDFQ-AV-----DIEAI
10ZF  -KA-GVALVTSVGPSCNLTITGMATANSEGDVVALGGAVKRADKAK--QV-HQ---SM-----DTVAM
2C31  -KP-GVCLTVSAPGFLNGVTSLAHATTNCFPMILLSGSSEREI--VDLQ--GDYE-EM-----DQMN
2AG0  -KL-GVALVTAGGGFTNAVTPIANAWLDRTPVFLFTGSGALRD--D--ET--NTLQAGI-----DQVAM
1V5F  -NL-GVTVSGSGPGASHLINGLYDAAMDNI PVVALTGSRPQRE--L--NM--DAFQ-EL-----NQNP
3EYA  -EL-AVCAGSCGPNLHLINGLFDCHRNHVPVLAIAAHIPSS--I--GS--GYFQ-ET-----HPQEL

2JLC  FASHPTHSISLPRPTQDIPARWL VSTIDHALGTL--H-AGGVHINCPFAEP-LYMDDTG-LS--WQQL-
1ZPD  AKNITAAEAAYTP-----EEAPAKIDHVIKTALRE-KKPVYLEIACN--IAS----M-PC-----A-
2VBF  HEPVTAARTLLTA-----ENATYEIDRVLSQLLKE-RKPVYINLPVD--VAA---A-KA-----E-
10VM  SENITVAQAVLTE-----QACYEIDRVL TMLRE-RRPGYLM L PAD--VAK---K-AA-----T-
2VK8  SANISETTAMITDI-----ATAPAEIDRCIRTYVT--QRPVYLG L PAN--LVD----L-NV--P---A-
2IHT  VAP-SKYAVELQRP-----HEITDLVDSAVNA--TEPVGSPFISLPVD--LLG---SSEG--I--D-
1BFD  PRPLVKWSYEPASA-----AEVPHAMSRATHMASMAPQGPVYLSVPYD--DWD---K-DA-DP---Q-
2PGN  FTPIARSTQRVERL-----DKVGEAIEAFRVAEGHPAGPAYVDIPFD--LTA---D-QI--DD-KA-
1JSC  SRSTKWNVMKSV-----EELPLRINEAFIATSGRPGPVLVDLPKD--VTA---A-ILRNP-----I-
1YBH  TRSITKHNYLVMDV-----EDIPRIIEEAFLATSGRPGPVLVDVPKD--IQQ---Q-LA--I---P-
2PAN  AKPVSK-AVTVREA-----ALVPRVLQQA FHL-RSGRPGPVLVDLPFD--VQV---A-EI--E---F-
10ZF  FSPVTKYAEVTAP-----DALAEVVSNAFRAAEQGRPGSAFVSLPQD--VVD---G-PV--S---G-
2C31  ARPHCKASFRINSI-----KDIPIGIARAVRTAVSGRPGGVYVDLPK--LFG---Q-TI--SVEEAN
2AG0  AAPITKWAHRVMAT-----EHIPRLVMQAIRAALSAPRGPVLLDLPWD--ILM---N-QI--DEDSV-
1V5F  YDHIAYNRRVAYA-----EQLPKLVDEAARMAIAK-RGVAVLEVPGD--FAK---V-EI--DNDQW-
3EYA  FRECSHYCELVSPP-----EQIPQVLAIAMRKA VLN-RGVSVVVLP GD--VAL---K-PA-PE--GA-

2JLC  -GDWWQDDKPWL-R--E----A-P---R--L-ES----EK--QRDWF FWRQ-K-RGVVAGR-MSA--E-
1ZPD  -A--P---G--P-A-SA----LFN---D--E-ASDEASLNAAVDETLKFIANRDKVAVLVGSKLR-AAG-
2VBF  -K--P---A--L-S--L-----E---N---TTEQVILSKIEESLKNAQKPVVIAGHEVI-SFG-
10VM  -P--P---V--N-A--L----THK---Q--A-HADSACLKAFRDAENKLAMSKRTALLADFLVL-RHG-
2VK8  -K--L--LQ--T-P--I---D-M---S-LK-PNDAESEKEVIDTILVLDKDAKNPVILADACCS-RHD-
2IHT  -P--N---P--P-A--N---T-PAKPV--G-VVA-DGWQKAADQAAALLAEAKHPVLVVGAAAI-RSG-
1BFD  -S--H---HLFD-R--H----V-S---S--S-VV--LNDQDLIDLKALNSASNPAIVLGPVDV-AAN-
2PGN  -L--V---P-RG-A--T---R-A---K--SVLH---APNEDVREAAAQLVAAKNPVILAGGVA-RSG-
1JSC  -P-----TKTTL-PSA-----QD--EFVMQSINKAADLINLAKKPVLYVGAGLL-NHAD
1YBH  -N--W--EQ--A-M--RLPGY--M---SRMP-KP---PEDSHLEQIVRLISESKKPVLYVGGGCL---N-
2PAN  -D--P--DY--E-P--L----P-V---Y--K-PA---ASR-QIEKAVEL-IQAERPVI VAGGGVI-NAD-

```

10ZF -K--V---L--P-----A-P---Q--M-GA---APDDAIDQVAKLIAQAKNPIFLGLMAS-QPE-
2C31 KL--L---F--K-P--I---DPA---P--A-QI---PAEDAIARAADLIKNAKRPVIMLGKGA- YAQ-
2AG0 -I--I---P--D-L--V---L-S---A--HGAR---PDPADLDQALALLRKAERPVIIVLGSSEAS-RTA-
1V5F -Y--S---S--ANS--L---R-K---Y--APIA---PAAQDIDAARELLNNSKRVPVIYAGIGTM---G-
3EYA -T--M---H--W-Y--H---A-P---Q--PVVT---PEEEELRKLQQLRYSSNIALMCGSGCA---G-

2JLC EGKKVALWAQTLGWPLIGD--VL-SQ-T-GQ-----PLPCA-----D--LWL-GNAKATSELO
1ZPD AEEAAVKFTDALGGAVATM--AAKSFF-PEE-NALYIGTSWGE---V-S--Y--PGV-EK--T---MK
2VBF LEKTVTQFVSETKLPITTL--NFGKSAV-DES-LPSFLGIYNGK---L-S--E--ISL-KN--F---VE
10VM LKHALQKVVKEVPMAHATM--LMGKGF-DER-QAGFYGTYSGS---A-S--T--GAV-KE--A---IE
2VK8 VKAETKKLIDLQFPFVVT--PMGKGS-DEQ-HPRYGGVYVGT---L-S--K--PEV-KE--A---VE
2IHT AVPAIRALAEARNIPVITT--YIAGV-L-PVG-HELNYGAVT-G---Y-DGILNFPA--LQ--T---FA
1BFD ANADCVMLAERLKAPVWVAPSAPR-CPF-PTR-HPCFRGLMP-AGIAA-----ISQL-----LE
2PGN GSEALLKLAEMVGVVVT--STGAGVF-PET-HALAMGSAG-F---C-G--W--KSA-ND--M---MA
1JSC GPRLLKELSDRAQIPVTTT--LQGLGSF-DQE-DPKSLDMLG-M---H-G--C--ATA-NL--A---VQ
1YBH SSDELGRFVELTGIPVAST--LMGLGSY-PD--DELSLHMLG-M---H-G--T--VYA-NY--A---VE
2PAN AAALLQFAELTSPVPIPT--L-GWGC-IPDD-HEL-AG-VG-L---QTA--H--RYG-NA--T---LL
10ZF NSKALRRLLETSHIPVTST--YQAAGAV-NQDNFSRFAGRVG-L---F-N-NQ--AG-DR--L---LQ
2C31 CDDEIRALVEETGIPFLPM--GMAKGLL-PDN-HPQSA-----A---TR-AF--A---LA
2AG0 RKTALSAFVAATGVPVFAD--YEGLSMLSG-LPDMRGLV--Q-----N---L-YS--F---A
1V5F HGPAVQELARKIKAPVITT--GKNFETF-EWD-FEALTGSTY-R---V-G--W--KPA-NE--T---IL
3EYA AHKELVEFAGKIKAPIVHA--LRGKEHV-EYD-NPYDVGMTG-L---I-G--F--SSG-FH--T---MM

2JLC ----QAQIVVQLGSSL---TGKRL-----QWQASC---EP-----EYWIWDDIEGRLD-PAHHR
1ZPD ----EADAVIALAPVF-NDYST--T--GW-TD--IP---DP-----KK-LVLAEP-----V
2VBF ----SADFLMLGVKL-TDSST--G--AF-TH--HL---DE-----NK-MISLNIDE-----GI
10VM ----GADTVLCVGRF-TDTLT--A--GF-TH--QL---TP-----AQ-TIEVQPHA-----AR
2VK8 ----SADLLSVGALL-SDFNT--G--SF-SY--SY--KTK-----N-IVEFHSDH-----MK
2IHT ----PVDLVLTVGYDYAED-LR-----P-S---WQKGIK-----K-TVRISPTVNPPIRV--YR
1BFD ----GHDVVLVIGAPVFRYH---QYDP--GQ--YL-KPGT-----R-LISVTCDPLEAARA--PM
2PGN ----AADFVLVLGSR-LSDWGI-AQ--GY-IT--KM---P-----K-FVHVDTPAVLGT--YF
1JSC ----NADLIIAVGARF-DDRVT--G--NI-SK--FA--PEARAAAEGRG-IIIHFVSPKNINKV--VQ
1YBH ----HSDLLAFGVRF-DDRVT--G--KL-EA--FA--SRA-----K-IVHIDISAEIGKN--KT
2PAN ----ASD-VFGIGNRF-ANRHT--G--SV-EK--YT--EGR-----K-IVHIDIEPTQIGRV--LC
10ZF ----LADLVICIGYSP-V--EY--E--P-AM--WN-SGNA-----T-LVHIDVLP-AYE-ERNYT
2C31 ----QCDVCLIGARL-NW-LM--Q--HG-KG-KTWGDELK-----K-YVQIDIQANEMDSN--QP
2AG0 KADAAPDLVLMGARF-GL-NT--G--HG-SG-QLI-PHSA-----Q-VIQVDPACELGRL--QG
1V5F ----EADTVLFAAGSNF-P---F--S-EVEG---TF-RNVD-----N-FIQIDPAMLGKR--HH
3EYA ----NADTLVLLGTQF-P---Y--R-----A--FY-PTDA-----K-IIQIDINPASIGAH--SK

2JLC --GRRLI-ANIADWLELH--PA--E-----KR-Q---P-W-CVEIPRLAEQAMQAV-IARR-D--A----
1ZPD VNGIRFPSVHLKDYLTRLAQKV---SK--K---TGLDFFKSLNAG--E-LK-K--AAPA-DPS--
2VBF IFNKVVEDFDRAVVSSSEL-K-G---IE--Y--E-----GQY-I-D--K-QYEE--F-IP-S-S--
10VM VGDVWFTGIPMNQAIETLVELCK-Q---HV--HA-P-----D---G--
2VK8 IRNATFPQVQMKFVLQKLLTTIA-D---AAKG-Y---K-P--VAVP--AR-TPANA-A---VP-A-S--
2IHT P-DVDVV-TDVLAFVEHFETATA-S-----FG-A---K-Q-R-HD--I-EPLRARI-AEFLADP-E-T-Y
1BFD G--DAIV-ADIGAMASALANLV-E-----ES-S---R-Q-L-PTA--A--PE-P-A-KV-D-A--
2PGN P--LLSV-ADAKTFMEQLIEVLP-GT--SGFKAURYQERENFRQATEFRAAWDGWV-REQE-SG-D-G--
1JSC T-QIAVE-GDATTNLGKMMSKI--F---PVK--E---RSEWFAQINKWKKEYP-YA-Y-ME-ET-P-G--
1YBH P-HVSV-CGDVKLALQGMNKVLENRA-EELKL-D---FGVWRNELNVQKQKFP-LS-F-K--TF-G-E--
2PAN P-DLGIV-SDAKAALTLVEVAQ-EQKAGRLP-C---RKEWVADCOQRKRTLL-RK-T--H-FD-N-V--
10ZF P-DVELV-GDIAGTLNKLQNI--D---HRL-V---LSPQAAEILDRQHQRELL-D-RRGAQLN-Q--
2C31 I-AAPVV-GDIKSAVSLLRKALK-G---AP-K--ADAEWGALKAKVDGNKAKL-AGKMTAE-T-PSG
2AG0 I-ALGIV-ADVGGTIEALAQATAQD---AAWP-D---RGDWCAKVTDLAQERYASI-AAKS-SS-E-H--
1V5F A-DVAIL-GDAALAIIDEILNKV--D-----AV-E---ESAWWTANLKNIANWREYI-NMLETKE-E-G--
3EYA V-DMALV-GDIKSTLRALLPLV--E-----EK-A---DRKFLDKALEYRDARKGL-DDLA-KP-S-E--

2JLC --FGEAQLAHRICDYL---P-E-QGQLFVGNS-LVVRLIDALS--QLPAG-YP-VYSNRGASGIDGLLST
1ZPD APLVNAEIAHQVEALL---T-P-NTTVIAETG-DSWFNAQRM---KLPNG-AR-VEYEMQWGHIGWSVPA
2VBF APLSQDRWLQAVESLT---Q-S-NETIVAEQG-TSFFGASTI---FLKSN-SR-FIQPLWGSIGYTFA
10VM -SLTQENFWRTLQTFI---R-P-GDIILADQG-TSAFGAIDL---RLPAD-VN-FIVQPLWGSIGYTLAA
2VK8 TPLKQEWMMNQLGNFL---Q-E-GDVVIAETG-TSAFGINQT---TFPNN-TY-GISQVLWGSIGFTGA
2IHT EDGRVHQVIDS-NTVEE-AAEPGEGTIVSDIG-FFRHYGVLF--RA-DQ-PFGFLTSAAGSSFGYGIPA
1BFD GRLHPETVFDLNDMA---P-E-NAIYLNST-STTAQMWRQL--NM-RN-PGSYYFA-AGGLGFALPA
2PGN MPASMFRAAEVRKVQ---R-P-EDIIVTDIG-NHTLPMFGGA--IL-QR-PRRLVTSMAEGILGCGFPM
1JSC SKIKPQTVIKKLSKVANDTG-R-HVIVTTGVG-QHQMWAAQHW--TW-RN-PHTFITSGLGTMGYGLPA
1YBH -AIPPQYAIKVLDEL---D-G-KAIISTGVG-QHQMWAAQFY--NY-KK-PRQLWSSGLGSMGVGLPA
2PAN -PVKQPVYEE-NKAF---G-R-DVCYVTTIG-LSQIAAAQL---HV-FK-DRHWINCGQAGPLGWTIPA
10ZF FALHPLRIVRAMQDIV---N-S-DVTLTVDMG-SFHIWIARYLYTFR-AR--Q-VMISNGQQTMGVALPW
2C31 -MMNYSNSLGVVRDFML-AN-P-DISLVNAGA-NALDNTRMIV--DM-LK-PRKRLDSGTWGMVGMIGMY
2AG0 -ALHPFHASQVIKHY---D-A-GVTVVADGA-LTYLWLSEVMSRVK-PG--G-FLCHGYLGSMTGVGFGT
1V5F -DLQFYQVYNAINNHA---D-E-DAIYSIDVGNSTQT-SIRHL--HM-TPKNM-WRTSPLFATMGIAIPG
3EYA KAIHPQYLAQQISHFA---A-D-DAIFTCDVG-TPTVWAARYL--KM-NGKRR-LLGSFNHGSANAMPO

A.4 A standard numbering scheme for thiamine diphosphate-dependent decarboxylases

2JLC AAGVQRAS----G--KPTLAIVGDL SALYDLNALALLR-QVSAPLVLIVVNN--GGQI-FSLLP--T-P
1ZPD AFGYAVGA----P-ERRNILMVG DGSFQLTAQEVAQMVRLLK-PVIFLINNY--GYTI-EVMIH----D
2VBF ALGSQIAD---K-ESRHLLFIGDGS LQLTVQELGLSIREKL-NPICFIINND--GYTV-EREIH----G
10VM AFGAQTAC----P-NRRVIVLTGDGAAQLTIQELGSMRLDKQ-HPITLVLNNE--GYTV-ERATH----G
2VK8 TLGAAFAAAEIDP-KKRVILFIGDGS LQLTVQEISTMIRWGL-KPYLFVLNND--GYTI-QKLIH----G
2IHT AIGAQ-AR----P-DQPTFLIAGDGGFHSNSSDLETTIARLNL-PIVTVVVVND--TNGL-IELYQNI-G
1BFD AIGVQLAE---P-ERQVIAVIGDGS ANYSISALWTAQAQYNI-PTIFVIMNNG--TYGA-LRWFAGVLEA
2PGN ALGAQLAE---P-NSRVFLGTGDGALYYHFNEFRVAVEHKL-PVITMVFTNE--SYGA-NWTLMNHQ-F
1JSC AIGAQVAK----P-ESLVIDIDGDASFNMTLTELSSAVOAGT-PVKILILNNEES-----
1YBH AIGASVAN---P-DAIVVDIDGDGS FIMNVQELATIRVENL-PVKVLLLNQ--HLGM-VMQWEDRF-Y
2PAN ALGVCAAD---P-KRNVVAISGDFDFQLIEELAVGAQFNI-PYIHVLVNN--YLGL-IRQSQ-RA-F
10ZF AIGAWLVN---P-ERKVVSVSGDGGF LQSSMELETAVRLKA-NVLHLIWDN--GYNM-VAIQEKK-Y
2C31 CVAAAAVT----G--KPVIAVEGDSAFGFSGMELETICRYNL-PVTVIIMNNG--GIYKG-NE-----
2AG0 ALGAQVAD---LEAGRRTILVTGDGS VGSIGEFDTLVRKQL-PLIVIIIMNNG--SWGA-TLHFQQLA-V
1V5F GLGAKNTY---P-DRQVWNIIGDGA FSMTPDVVTNVRYNM-PVINVVSNT--EYAF-IKNKYEDTNK
3EYA ALGAQATE---P-ERQVVAMCGDGGFS MLMGDFLSVVQMKL-PVKIVVFNS--VLGF----V-----

2JLC -Q-SERE-----R-FYLMQN-----VHFEHAAAMF-----E-LKYHRPQ-NWQLETA
1ZPD -----G-----P-Y-N-NIK-----NWDYAGLMEVFNNGGYDSGAAGLKK-KTGGELAEA
2VBF -P-T-Q-----S-Y-N-DIP-----MWNYSKLPETFGA----TEDRVVSKIVR-TENEFVSV
10VM -A-E-Q-----R-Y-N-DIA-----LWNWTHIPQALS-----LDPQSECWRVS-EAEQLADV
2VK8 -P-K-A-----Q-Y-N-EIQ-----GWDHLSLLPTFG-----AKD-YETHRVA-TTGEWDKL
2IHT HR-S-H-----DPA-V-KFG-----GVDFVALAEAN-----G-VDATRAT-NREELLAA
1BFD E--N-V-----P-G-L-DVP-----GIDFRALAKGY-----G-VQALKAD-NLEQLKGS
2PGN G--Q-N-----N-W-T-EFM-----NPDWVGIAKAF-----G-AYGESVRETG-DIAGA
1JSC -----H-----T-HQL-----NPDFIKLAEAM-----G-LKGLRVK-KQEELDAK
1YBH KA-N-R-----A-H-T-FLGDPAQ---EDEFNMLLFAAAC-----G-IPAARVT-KKADLREA
2PAN D-----D-----Y-C-V-QLAFENINSSEVNGYVDHVKVAEGL-----G-CKAIRVF-KPEDIAPA
10ZF -Q-R-L-----S-G-V-EFG-----PMDFKAYAESF-----G-AKGF AVE-SAEALEPT
2C31 --AD--PQPGVIS-C-T-RLT-----RGRYDMMMEAF-----G-GKGYVAN-TPAELKAA
2AG0 GPNR--V-----T-G-T-RLE-----NGSYHGVAAAF-----G-ADGYHVD-SVESFSAA
1V5F NL----F-----G--V-DFT-----DVDYAKIAEAQ-----G-AKGFTVS-RIEDMDRV
3EYA -----G--T-ELH-----DTNFARIAEAC-----G-ITGIRVE-KASEVDEA

2JLC FAD--AW-RT---P-T-TTVIEMVVNDTD-G--AQTL--Q-----QLLAQVSHL-----
1ZPD IKV--AL-AN---T-DGPTLIECFIGRED-C--T--EELVKWGRVAAAANSRKPVNK-----
2VBF MKE--AQADV---N-R-MYWIELVLEKED-A--P--KLLKKMGKLF AEQNK-----
10VM LEK--VA-HH---E-R-LSLIEVMLPKAD-I--P--PLL GALTKALEACNN-----
2VK8 TQDKSFN-DN---S-K-IRMIEVMLPVFD-A--P--Q-NLV--EQAK--LTAATNAKQ-----
2IHT LRK--GA-EL---G-R-PFLIEVPVN-YD-F--Q--PGG--FGALSI-----
1BFD LQE--AL-SA---K-G-PVLIEVSTV-----
2PGN LQR--AI-DS---G-K-PALIEIPVSKTQ-GLAS--D--PVG-GVGNLLLKGREIPVDTTGSMYPGENL
1JSC LKE--FV-ST---K-G-PVLLEVEVDKK-----
1YBH IQT--ML-DT---P-G-PYLLDVICPHQE-H--V--L--P--MIPSGGTFNDVITEGDGR-----
2PAN FEQ--AK-AL--AQYRVPVVVEVILERV-T-N--I-----S--GSELDNVEFEDIADNAADAPTETCFH
10ZF LRA--AM-DV---D-G-PAVVAIPVDYRDNP-LL---MGQ--LH-----
2C31 LEE--AV-AS---G-K-PCLINAMIDPDAGV-----E-----
2AG0 LAQ--AL-AH---N-R-PACINVAVALDP-I--P--PEEL----I-----
1V5F MAE--AV-AANKAG-H-TVVIDCKIT-QD-R--P--I--PV---ETLKLDSKLYSEDEIKAYKERYEAAAN
3EYA LQR--AF-SI---D-G-PVLVDVVVA-KE-E---L--A---IPPQIK-----

2JLC -----
1ZPD -----
2VBF -----
10VM -----
2VK8 -----
2IHT -----
1BFD -----
2PGN LHLK-----
1JSC -----
1YBH -----
2PAN YE-----
10ZF -----
2C31 -----
2AG0 -----
1V5F LVPFREYLEAEGLESKYIK
3EYA -----

A.5 The modular structure of ThDP-dependent enzymes

Table A.5: Seed sequences chosen for the generation of the Thiamine diphosphate-dependent Enzyme Engineering Database. Eventually appended His-tags were removed from the respective sequences.

Protein name	Source organism	NCBI accession code	PDB accession code (chain identifier)
Pyruvate oxidase	<i>Lactobacillus plantarum</i>	gi 494458	pdb 1POX (A)
Pyruvate oxidase	<i>Escherichia coli</i>	gi 211939453	pdb 3EYA (A)
Indolepyruvate decarboxylase	<i>Enterobacter cloacae</i>	gi 31615929	pdb 1OVM (A)
Phenylpyruvate decarboxylase	<i>Komagataella pastoris</i> GS115	gi 254570389	
Pyruvate decarboxylase E477Q variant	<i>Saccharomyces cerevisiae</i>	gi 222142974	pdb 2VK8 (A)
Pyruvate decarboxylase	<i>Acetobacter pasteurianus</i>	gi 178847304	pdb 2VBI (A)
Branched-chain keto acid decarboxylase	<i>Lactococcus lactis</i>	gi 161172287	pdb 2VBF (A)
2-Succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexadiene-1-carboxylate synthase	<i>Escherichia coli</i>	gi 209870258	pdb 2JLC (A)
Menaquinone biosynthesis protein MenD	<i>Bacillus subtilis</i>	gi 300508355	pdb 2X7J (A)
Chloroplast Phyllo enzyme	<i>Arabidopsis thaliana</i>	gi 71738181	
Benzoylformate decarboxylase	<i>Pseudomonas putida</i>	gi 157830273	pdb 1BFD (A)
Benzaldehyde lyase	<i>Pseudomonas fluorescens</i>	gi 88192045	pdb 2AG0 (A)
Oxalyl-Coa decarboxylase	<i>Escherichia coli</i>	gi 189339519	pdb 2Q28 (A)
2-Hydroxyphytanoyl-CoA lyase	<i>Mus musculus</i>	gi 6688685	
Acetolactate synthase I	<i>Escherichia coli</i> str. K-12 substr. MG1655	gi 16131541	
Acetohydroxyacid synthase	<i>Saccharomyces cerevisiae</i>	gi 18655870	pdb 1JSC (A)
Acetolactate synthase	<i>Geobacillus thermoglucosidasius</i> C56-YS93	gi 336236810	
Acetolactate synthase	<i>Klebsiella pneumoniae</i>	gi 38492715	pdb 1OZG (A)
Sulfoacetaldehyde acetyltransferase	<i>Burkholderia phymatum</i> STM815	gi 186470983	
Carboxyethylarginine synthase	<i>Streptomyces clavuligerus</i>	gi 39654981	pdb 1UPA (A)
Glyoxylate carbonylase	<i>Escherichia coli</i>	gi 164414776	pdb 2PAN (A)
CDP-Yersiniose biosynthesis protein	<i>Yersinia pseudotuberculosis</i>	gi 347810936	
Cyclohexane-1,2-dione hydrolase	<i>Azoarcus sp.</i>	gi 185177534	pdb 2PGN (A)

continued Table A.5

Protein name	Source organism	NCBI accession code	PDB accession code (chain identifier)
3D-(3,5/4)-Trihydroxycyclohexane-1,2-dione hydrolase	<i>Pseudomonas synxantha</i>	gi 492245762	
PigD	<i>Serratia marcescens</i> WW4	gi 448240871	
Transketolase	<i>Saccharomyces cerevisiae</i>	gi 20149857	pdb 1GPU (A)
Transketolase	<i>Homo sapiens</i>	gi 312208042	pdb 3OOY (A)
Phosphoketolase	<i>Bifidobacterium breve</i>	gi 304445733	pdb 3AHC (A)
Dihydroxyacetone synthase	<i>Aspergillus oryzae</i> RIB40	gi 169768774	
1-Deoxy-D-xylulose 5-phosphate synthase	<i>Escherichia coli</i>	gi 122921297	pdb 2O1S (A)
Pyruvate:ferredoxin oxidoreductase	<i>Desulfovibrio africanus</i>	gi 90108959	pdb 2C3P (A)
Pyruvate:flavodoxin/ferredoxin oxidoreductase domain-containing protein	<i>Methanohalobium evestigatum</i> Z-7303	gi 298675714	
Pyruvate:ferredoxin oxidoreductase subunit beta	<i>Methanocella conradii</i> HZ254	gi 383319282	
hypothetical protein	<i>Pyrococcus furiosus</i>	gi 61680602	pdb 1YD7 (A)
2-Oxoglutarate:ferredoxin oxidoreductase beta subunit	<i>Azospirillum brasilense</i> Sp245	gi 392381025	
Indolepyruvate:ferredoxin oxidoreductase subunit alpha	<i>Pyrococcus yayanosii</i> CH1	gi 337284193	
Indolepyruvate:ferredoxin oxidoreductase	<i>Janthinobacterium</i> sp. HH01	gi 495719244	
Pyruvate dehydrogenase	<i>Escherichia coli</i>	gi 116668117	pdb 2IEA (A)
Branched-chain dehydrogenase (E1)	<i>Pseudomonas putida</i>	gi 75766368	pdb 2BP7 (A)
Branched-chain dehydrogenase (E1)	<i>Pseudomonas putida</i>	gi 75766369	pdb 2BP7 (B)
Pyruvate dehydrogenase	<i>Pyrobaculum aerophilum</i>	gi 18158937	pdb 1IK6 (A)
Pyruvate dehydrogenase (E1p)	<i>Homo sapiens</i>	gi 215261326	pdb 3EXE (A)
Pyruvate dehydrogenase (E1p)	<i>Homo sapiens</i>	gi 215261327	pdb 3EXE (B)
Branched-chain 2-oxo acid dehydrogenase (E1)	<i>Thermus thermophilus</i>	gi 47169251	pdb 1UMB (A)
Branched-chain 2-oxo acid dehydrogenase (E1)	<i>Thermus thermophilus</i>	gi 47169252	pdb 1UMB (B)
Sulfopyruvate decarboxylase subunit alpha	<i>Pseudomonas</i> sp. UW4	gi 426410237	
Sulfopyruvate decarboxylase	<i>Pseudomonas</i> sp. Lz4W	gi 459961266	

continued Table A.5

Protein name	Source organism	NCBI accession code	PDB accession code (chain identifier)
Phosphonopyruvate decarboxylase	<i>Agrobacterium vitis</i> S4	gi 222149501	
Phosphonopyruvate decarboxylase	<i>Azoarcus</i> sp. BH72	gi 119898989	
2-Oxoglutarate dehydrogenase (E1o)	<i>Escherichia coli</i> K-12	gi 134104925	pdb 2JGD (A)
α -Ketoglutarate decarboxylase	<i>Mycobacterium smegmatis</i> str. MC2 155	gi 340707373	pdb 2XT6 (A)

Table A.6: Location of the PP and PYR domains in known structures of ThDP-dependent enzymes. The start and end positions given in this table are the absolute positions relating to the start and end positions of the respective domain in the respective structure. Empty fields indicate missing structural information about the respective domains. The letters in brackets indicate the respective protein chain.

PDB	PP	PYR	PDB	PP	PYR
1AY0	30-248 (A)	358-527 (A)	2Q28	372-542 (A)	10-167 (A)
1B0P	817-1070 (A)	10-172 (A)	2QTA	108-391 (A)	496-692 (A)
1DTW	92-285 (A)		2R5N	27-245 (A)	357-519 (A)
1GPU	30-248 (A)	358-527 (A)	2R8O	27-245 (A)	357-519 (A)
1IK6		10-177 (A)	2UZ1	372-544 (A)	6-162 (A)
1ITZ	39-259 (A)	368-532 (A)	2V3W	355-523 (A)	4-159 (A)
1L8A	106-391 (A)	496-692 (A)	2VBF	356-526 (A)	4-165 (A)
1NI4	68-257 (A)	5-172 (B)	2VBI	363-531 (A)	5-167 (A)
1OLX		23-191 (B)	2VK1	367-540 (A)	6-168 (A)
1OVM	362-530 (A)	7-168 (A)	2VK4	367-540 (A)	6-168 (A)
1OZH	373-541 (A)	13-168 (A)	2VK8	367-540 (A)	6-168 (A)
1PI3	355-523 (A)	4-159 (A)	2W93	367-540 (A)	6-168 (A)
1PO7	355-523 (A)	4-159 (A)	2WVG	367-539 (A)	5-167 (A)
1POW	372-544 (A)	13-169 (A)	2X7J	383-554 (A)	5-172 (A)
1POX	372-544 (A)	13-169 (A)	2XT6	511-741 (A)	877-1069 (A)
1PYD	367-540 (A)	6-168 (A)	2YIC	511-741 (A)	877-1069 (A)
1Q6Z	355-523 (A)	4-159 (A)	3AHC	67-298 (A)	411-604 (A)
1QGD	27-245 (A)	357-519 (A)	3AHG	67-298 (A)	411-604 (A)
1QPB	367-540 (A)	6-168 (A)	3AHH	67-298 (A)	411-604 (A)
1QS0	112-306 (A)	10-175 (B)	3AHI	67-298 (A)	411-604 (A)
1R9J	26-246 (A)	352-517 (A)	3AHJ	67-298 (A)	411-604 (A)
1T9B	473-645 (A)	93-250 (A)	3AI7	67-299 (A)	411-604 (A)
1TKB	31-248 (A)	358-527 (A)	3D7K	372-544 (A)	6-162 (A)
1UMD	72-267 (A)	6-173 (B)	3DUF	82-265 (A)	5-172 (B)
1UPA	383-559 (A)	14-169 (A)	3DVA	82-265 (A)	5-172 (B)
1V11	92-285 (A)	24-190 (B)	3E9Y	464-641 (A)	99-255 (A)
1V16	92-285 (A)	24-190 (B)	3EA4	464-641 (A)	99-255 (A)
1V1M	92-285 (A)	24-190 (B)	3EXE	68-257 (A)	5-172 (B)
1V1R	92-285 (A)	24-190 (B)	3EXH	68-257 (A)	5-172 (B)
1V5F	367-539 (A)	8-164 (A)	3EXI	68-258 (A)	5-172 (B)
1V5G	367-539 (A)	8-164 (A)	3EYA	359-527 (A)	5-160 (A)
1W88	81-265 (A)	5-172 (B)	3F6E	359-523 (X)	4-159 (X)
1X7W	92-285 (A)	22-190 (B)	3FLM	368-537 (A)	2-169 (A)
1X7X	92-285 (A)	22-190 (B)	3FSJ	355-523 (X)	4-159 (X)
1X7Y	92-285 (A)	22-190 (B)	3HWW	369-537 (A)	3-169 (A)
1X7Z	92-285 (A)	22-190 (B)	3HYL	32-247 (A)	357-520 (A)
1X80	92-285 (A)	22-190 (B)	3IAE	372-544 (A)	6-162 (A)
1YBH	464-641 (A)	99-255 (A)	3IAF	372-544 (A)	6-162 (A)
1YD7		17-178 (A)	3KOM	29-247 (A)	357-520 (A)
1YHY	464-641 (A)	99-255 (A)	3L84	29-240 (A)	340-498 (A)
1YHZ	464-641 (A)	99-255 (A)	3LPL	107-390 (A)	496-692 (A)
1YI0	464-641 (A)	99-255 (A)	3LQ1	386-555 (A)	4-174 (A)
1YI1	464-641 (A)	99-255 (A)	3LQ2	107-391 (A)	496-692 (A)

1YNO	355-523 (A)	4-159 (A)	3M49	35-247 (A)	357-520 (A)
1Z8N	464-641 (A)	99-255 (A)	3M6L	29-240 (A)	340-498 (A)
1ZPD	367-539 (A)	5-167 (A)	3OE1	367-539 (A)	5-167 (A)
2AG0	372-544 (A)	6-162 (A)	3OOY	38-242 (A)	317-473 (A)
2AG1	372-544 (A)	6-162 (A)	3RIM	46-267 (A)	377-551 (A)
2BEW	92-285 (A)	24-190 (B)	3UK1	47-264 (A)	376-543 (A)
2BFB	92-285 (A)	24-190 (B)	3UPT	47-264 (A)	376-543 (A)
2BFC	92-285 (A)	24-190 (B)	3ZHR	511-741 (A)	877-1069 (A)
2BFD	92-285 (A)	24-190 (B)	4GG1	355-523 (A)	4-159 (A)
2BFE	92-285 (A)	24-190 (B)	4GM0	355-523 (A)	4-159 (A)
2BFF	92-285 (A)	24-190 (B)	4GM1	355-523 (A)	4-159 (A)
2BP7	112-307 (A)	8-175 (B)	4GM4	355-523 (A)	4-159 (A)
2C31	377-546 (A)	13-169 (A)	4GP9	355-523 (A)	4-159 (A)
2C42	817-1070 (A)	10-172 (A)	4GPE	355-523 (A)	4-159 (A)
2E6K	31-249 (A)	351-514 (A)	4JD5	355-523 (A)	4-159 (A)
2EZ9	372-545 (A)	13-169 (A)	4JU8	355-523 (A)	4-159 (A)
2FN3	355-524 (A)	4-159 (A)	4JU9	355-523 (A)	4-159 (A)
2FWN	355-524 (A)	4-159 (A)	4JUA	355-523 (A)	4-159 (A)
2IEA	107-391 (A)	496-692 (A)	4JUB	355-523 (A)	4-159 (A)
2IHT	383-559 (A)	14-169 (A)	4JUC	355-523 (A)	4-159 (A)
2IHU	383-559 (A)	14-169 (A)	4JUD	355-523 (X)	4-159 (X)
2J9F	92-285 (A)	23-190 (B)	4JUF	355-523 (A)	4-159 (A)
2JGD	232-454 (A)	592-778 (A)	4K9K	355-523 (A)	4-159 (A)
2JLA	369-537 (A)	3-169 (A)	4K9L	355-523 (A)	4-159 (A)
2JLC	369-537 (A)	3-169 (A)	4K9M	355-523 (A)	4-159 (A)
2NXW	356-520 (A)	3-164 (A)	4K9N	355-523 (A)	4-159 (A)
2O1S	50-282 (A)	320-477 (A)	4K9O	355-523 (A)	4-159 (A)
2O1X	52-287 (A)	324-479 (A)	4K9P	355-523 (A)	4-159 (A)
2OZL	68-257 (A)	7-172 (B)	4KGD	372-544 (A)	13-169 (A)
2PAN	372-555 (A)	8-163 (A)	4KXV	38-242 (A)	317-473 (A)
2PGN	377-545 (A)	6-163 (A)			

Table A.7: Overview over the alteration of secondary structure elements in the PYR and PP domains of ThDP-dependent enzymes.

Region	PYR		PP		
1	TK (non-human)	loop extended	TK	PP- α B \rightarrow α -helix \rightarrow coil \rightarrow PP- β 2	
	OR	loop extended	DXPS	PP- α B \rightarrow α -helix \rightarrow coil \rightarrow PP- β 2	
	aKGDH	loop extended	aKGDH	PP- α B \rightarrow α -helix \rightarrow β -strand \rightarrow extended PP- β 2	
	aKADH1	loop extended	aKADH 1	PP- α B \rightarrow α -helix \rightarrow coil \rightarrow PP- β 2	
			aKADH 2	PP- α B \rightarrow α -helix \rightarrow coil \rightarrow PP- β 2	
2	aKGDH	insertion of short antiparallel β -sheet			
3	TK	replaced by short loop	TK	short antiparallel β -sheet	
	DXPS	replaced by short loop	DXPS	replaced by short loop; ~40 residues missing in structures	
	aKADH 1	replaced by short loop	aKADH 1	replaced by short loop	
	aKADH 2	replaced by short loop	aKADH 2	replaced by short loop	
	aKGDH	replaced by short loop	aKGDH	replaced by short loop	
	OR	replaced by short loop	OR	replaced by short loop	
4			OR	extended PP- α C plus 3 additional α -helices	
5			aKADH 1	6 short α -helices forming a small additional domain within the PP-domain	

A.6 A Tailor-Made Chimeric Thiamine Diphosphate-Dependent Enzyme for the Direct Asymmetric Synthesis of (S)-Benzoin

A.6.1 Sequence and structural analysis of *Ap*PDC and *Pf*BAL

Sequence analysis

The homologous families containing pyruvate decarboxylase (PDC) from *Acetobacter pasteurianus* and benzaldehyde lyase (BAL) from *Pseudomonas fluorescens* (*Pf*BAL) of an up-to-date version of the thiamine diphosphate (ThDP)-dependent enzyme engineering database (TEED) (Widmann, Radloff, and Pleiss 2010) were considered for conservation analysis (unpublished data). 186 sequences of *Ap*PDC homologues and 43 homologous *Pf*BAL sequences were aligned using the standard numbering approach for ThDP-dependent decarboxylases (Vogel, Widmann, et al. 2012) and subsequently analyzed regarding their amino acid distribution. The absolute position numbers of the *Ap*PDC sequence (Rother et al. 2011) and *Pf*BAL sequence (gi|9965497) were counted increasingly ordered beginning with methionine in position 1.

Structural analysis

The PyMOL software (Schrödinger) (*The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC, <http://pymol.org> 2013*) was used for structural investigations of *Ap*PDC (pdb|2VBI), *Ap*PDC variants, and *Pf*BAL (pdb|2AG0). The hydroxybenzyl-ThDP (HBThDP) intermediate was constructed using the ArgusLab software (Mark A. Thompson, Planaria Software LLC, Seattle, USA) and modeled into the active sites of the *Ap*PDC and *Pf*BAL structures. Both structures were then manually superimposed by aligning the ThDP molecules that were co-crystallized with the respective proteins. The YASARA software (YASARA Biosciences GmbH, Austria) was applied to equilibrate the structure of *Ap*PDC-E469G/T384G (see Publication, Figure 4.10B) using the YASARA2 force field (Krieger, Koraimann, and Vriend 2002). Therefore, the structure was embedded in a water box with 2 Å distance to the protein and prepared by applying the standard energy minimization protocol. A subsequent simulation run for 0.5 ns was performed in order to equilibrate the structure of the protein and the HBThDP

intermediate. A constraint was set to the HBThDP intermediate to keep the planar orientation of the phenyl ring of benzaldehyde, which is due to mesomeric stabilization in the physiological structure.

A.6.2 Conservation of aromatic amino acids in ThDP-dependent decarboxylases

A conserved cluster of aromatic residues was found in close spatial proximity to W543 in *Ap*PDC. Conservation analysis with ThDP-dependent decarboxylases (family classification according to Duggleby 2006 and Widmann, Radloff, and Pleiss 2010), based on an up-to-date version of the TEED (Widmann, Radloff, and Pleiss 2010) (unpublished data), revealed that this cluster is exclusively conserved in sequences homologous to *Ap*PDC (Table A.8 on page 219), which form the second homologous family of PDCs in the TEED (Widmann, Radloff, and Pleiss 2010).

A.6.3 Generation, expression and purification of *Ap*PDC variants

Generation of *Ap*PDC variants by site-directed mutagenesis

All variants were constructed as earlier described using the standard Quikchange[®] site-directed mutagenesis protocol (Stratagene) (Westphal, Hahn, et al. 2013). The plasmid pET22b (Novagen) containing the gene encoding *Ap*PDC-E469G (Rother et al. 2011) was used as template for the preparation of the mutant encoding *Ap*PDC-E469G/T384G, using the forward (fw) primer 5'-CGCTGGTGGCAGAAaggCGGCGATTCATGG-3' and the reverse primer 5'-CCATGAATCGCCGccTTCTGCCACCAGCG-3' (mutated codons are underlined, with lower-case letters indicating base changes). Based on the gene sequences encoding *Ap*PDC-E469G/T384G, respective mutants encoding *Ap*PDC-E469G/T384G/I468G, *Ap*PDC-E469G/T384G/I468A, and *Ap*PDC-E469G/T384G/I468V were prepared. The following fw and rv primers were used for mutagenesis. Respective mutated positions are underlined:

*Ap*PDC-E469G/T384G/I468G: 5'-CGTGGCTATGTCggcGGCATCGCCATTC-3' (fw)

5'-GAATGGCGATGCCgccGACATAGCCACG-3' (rv)

*Ap*PDC-E469G/T384G/I468A: 5'-CCGTGGCTATGTCgccGGCATCGCCATTC-3' (fw)

5'-GAATGGCGATGCCggcGACATAGCCACGG-3' (rv)
*Ap*PDC-E469G/T384G/I468V: 5'-CAACCGTGGCTATGTCgTgGGCATC
GCCATTCATG-3' (fw)
5'-CATGAATGGCGATGCCcAcGACATA
GCCACGGTTG-3' (rv)

After digestion of parental DNA by *DpnI*, chemically competent *E. coli* BL21 (DE3) cells were transformed by the resulting PCR products using the CaCl₂-method according to Sambrook and Russell (Sambrook and Russel 2001). Gene sequences were confirmed by DNA sequencing (LGC Genomics, Berlin, Germany).

Expression of *Ap*PDC variants

Cultivation of the variants was performed in shaking flasks (5 L) at initially 30 °C using 1 L of lysogeny broth (LB) medium, pH 7.0, supplemented with ampicillin (100 µg µL⁻¹). Expression of the *Ap*PDC constructs in the vector pET22b, which provides additional codons for a C-terminal 6 × histidine tag, was induced by addition of isopropyl-β-D-1-thiogalactopyranoside (0.4 mM) at an OD₆₀₀ of 0.6. After induction, cells were grown overnight at 20 °C, subsequently harvested by centrifugation (30 min, 15.000 g, 4 °C), and finally stored at -20 °C.

Purification of *Ap*PDC variants

For purification cells were resuspended in disintegration buffer (50 mM triethanolamine (TEA), 150 mM NaCl, 2 mM MgSO₄, 0.1 mM ThDP, 20 mM imidazole, pH 7.5) supplemented with lysozyme (1 mg mL⁻¹) on ice. After cell disruption by sonication, cell debris was removed by centrifugation (45 min, 48.000 g, 4 °C). His-tagged *Ap*PDC variants were then purified by immobilized nickel ion chelate chromatography: washing buffer (50 mM TEA, 150 mM NaCl, 50 mM imidazole, pH 7.5); elution buffer (50 mM TEA, 150 mM NaCl, 250 mM imidazole, pH 7.5). Subsequently, purified enzymes were desalted by size exclusion chromatography using Sephadex™ G25M (GE Healthcare), equilibrated with a low salt buffer (10 mM TEA, 2 mM MgSO₄, 0.1 mM ThDP, pH 7.5). Finally, *Ap*PDC variants were freeze-dried using the lyophilizer ALPHA 2-4 (Christ, Germany). Lyophilized *Ap*PDC variants were stored at -20 °C for several months without significant loss of activity.

Table A.8: Conservation of aromatic amino acids in ThDP-dependent decarboxylases at positions corresponding to the aromatic cluster observed in *ApPDC*.

HFAM ^b	Standard position ^a (absolute position in <i>ApPDC</i>)				
	264 (262)	292 (290)	392 (388)	393 (389)	559 (543)
Conserved aromatic amino acids (frequency in %)					
IPD	-	-	-	F (73), Y (19), W (1)	-
PhePDC	-	-	-	F (89), Y (4)	-
PDC ^c	-	F (78), Y (2)	-	F (88), Y (2)	-
PDC ^d	W (99)	Y (91), F (2)	W (98)	F (98)	W (76), F (4)
KdcA	-	-	-	F (51), Y (8)	-
MenD	W (63)	-	-	-	-
AHAS	-	-	-	-	W (78), F (6)
AHAS	Y (43), F (14)	-	Y (50)	-	-
AHAS	-	-	F (81)	-	-
AHAS	-	W (61)	-	-	-
SAAT	-	F (72), W (16), Y (3)	-	-	-
CDH	-	W (100)	-	-	-
THcHDPH	-	Y (100)	-	-	-

^a Standard numbers according to the standard numbering scheme for ThDP-dependent decarboxylases (Vogel, Widmann, et al. 2012).

^b Names of homologous families (HFAM): IPDC, indolepyruvate decarboxylases; PhePDC, phenylpyruvate decarboxylases; PDC, pyruvate decarboxylases; AHAS, acetohydroxyacid synthases; SAAT: sulfoacetaldehyde acetyltransferase; CDH, cyclohexane-1,2-dione hydrolase; THcHDOH: 3D-(3,5/4)-trihydroxycyclohexane-1,2-dione hydrolase. Only those homologous families with at least ten incorporated sequences and the CDH family were analyzed.

^c Sequences homologous to pyruvate-activated PDC from *Saccharomyces cerevisiae*.

^d Sequences homologous to the *ApPDC* and the PDC from *Zymomonas mobilis*.

A.6.4 Single site-saturation mutagenesis of position 543

For the generation of the mutant library for mutation of codon TGG encoding tryptophan at amino acid position 543, the gene coding for *ApPDCE469G/T384G/I468A* was used as template.

The fw primer 5'-GGATATGCTGGTTCAAAndtGGCCGCAAGGTTGCC-3' and

the rv primer 5'-GGCAACCTTGCGGCCahnTTGAACCAGCATATCC-3',

which contained NDT degeneracy (where N = any nucleotide, D = A, G, or T), were applied in the standard Quikchange[®] mutagenesis protocol (Stratagene). Library preparation, expression

of the variants, and cell disruption were performed as previously described (Westphal, Hahn, et al. 2013). For the carboligation assay in a 96 deep well plate (max. volume 2 mL; Nerbe plus, Germany), 200 μL of the respective supernatant was mixed with 300 μL of a benzaldehyde stock solution (30 mM benzaldehyde in reaction buffer; see below) giving a final concentration of 18 mM benzaldehyde. Samples of 50 μL and 100 μL of the reaction mixture were then transferred to two new 96 deep well plate, respectively. The plates were covered with a silicone cap-mat (Nerbe plus, Germany) and incubated at 20 °C and 600 rpm (VARIOMAG Shaker, H + P Labortechnik, Germany) for 24 h to allow biocatalytic formation of benzoin. Subsequently, the 100 μL batches were analyzed for benzoin formation via the colorimetric triphenyltetrazolium chloride assay as earlier described (Westphal, Hahn, et al. 2013; Breuer et al. 2002). In the case of positive hits, indicated by red coloration, the reaction mixture of the 50 μL batch was extracted with 500 μL methyl tert-butyl ether (MTBE) and the organic phase was analyzed by chiral-phase HPLC (see Section A.6.7). The resulting *Ap*PDC variant E469G/T384G/I468A/W543F with improved (*S*)-selectivity for the formation of benzoin was finally expressed and purified as described before (see Section A.6.3).

A.6.5 Carboligation reactions

All carboligation reactions were performed in reaction buffer (50 mM TEA, 2 mM MgSO_4 , 0.1 mM ThDP, pH 8.0, if not otherwise indicated) using lyophilized enzyme. The protein concentration was determined according to Bradford using bovine serum albumin as a standard (Bradford 1976).

Analytical scale

Carboligation reactions were performed in a total volume of 300 μL in glass vials at 800 rpm using a thermomixer (Eppendorf, Germany). The final protein concentration was adjusted to 1 mg mL^{-1} . The enzyme was incubated with benzaldehyde and benzaldehyde derivatives (18 mM), respectively, for the homocoupling reaction, or with acetaldehyde (18 mM) and benzaldehyde (18 mM) for the heterocoupling reaction. Subsequently, products were extracted using 300 μL of MTBE. Finally, the organic phase was used for product analysis (see Section A.6.7 on page 222). Reactions were performed in triplicate. Reaction conditions were varied as indicated. No C-C bond formation was observed in control experiments without enzyme.

Preparative scale

*Ap*PDC variants E469G/T384G/I468A and E469G/T384G/I468A/W543F (1 mg mL⁻¹ final concentration) were incubated in 20 mL reaction buffer (see above) with the respective benzaldehyde derivative (18 mM) in a 50 mL Falcon tube using a thermomixer (Eppendorf, Germany) at 15 °C and 750 rpm. The reaction was monitored for 24–48 h until conversions >90% were achieved. In case of *Ap*PDCE469G/T384G/I468A/W543F the reaction was supplemented with 1 mg mL⁻¹ fresh enzyme after 12 h and 24 h. The reaction mixtures were then extracted three times using dichloromethane (20 mL). Afterwards, the organic phase was washed with water and brine and dried over Na₂SO₄, as described earlier (Demir, Sesenoglu, Eren, et al. 2002). Finally, the solvent was evaporated and crystallization afforded the respective pure (*S*)-benzoin derivatives.

A.6.6 Reaction optimization

To further optimize the synthesis of (*S*)-benzoin by *Ap*PDC variants, the influence of pH, substrate concentration, temperature, and the addition of organic cosolvents on both conversion and stereoselectivity was investigated. Variation of pH and substrate concentration influenced the catalytic activity of both variants but not the stereoselectivity. Best conversion was obtained at pH 8.0, exemplarily shown for *Ap*PDC-E469G/T384G/I468A in Figure A.4.

Higher substrate concentrations (≥ 30 mM) led to fast precipitation of the enzyme. In contrast, temperature and the addition of cosolvents were identified to affect the stereoselectivity. Reduced temperature resulted in improved (*S*)-selectivity. For example, *Ap*PDC-E469G/T384G/I468A/W543F yielded (*S*)-benzoin with 93% ee and a conversion of 44% at 30 °C, whereas at 15 °C (*S*)-benzoin was formed with 98% ee, but accompanied by a reduced conversion of 26% (Figure A.5). In the presence of dimethyl sulfoxide (DMSO) or MTBE as cosolvents, both common additives to enhance the solubility of aromatic substrates, (*S*)-selectivity was significantly reduced. For example, *Ap*PDC-E469G/T384G/I468A-catalyzed formation of (*S*)-benzoin in the presence of DMSO (20 vol%) and MTBE (5 vol%) resulted in reduced ee-values of 75% and 80%, respectively, compared to 87% ee in the absence of cosolvents. In addition, conversion dropped from 95% without cosolvent to 28% and 42%, respectively, in the presence of the cosolvents. These results are consistent with recently published data (Gehards et al. 2012), where the influence of organic solvents on the stereoselectivity of ThDP-dependent enzymes was comprehensively investigated.

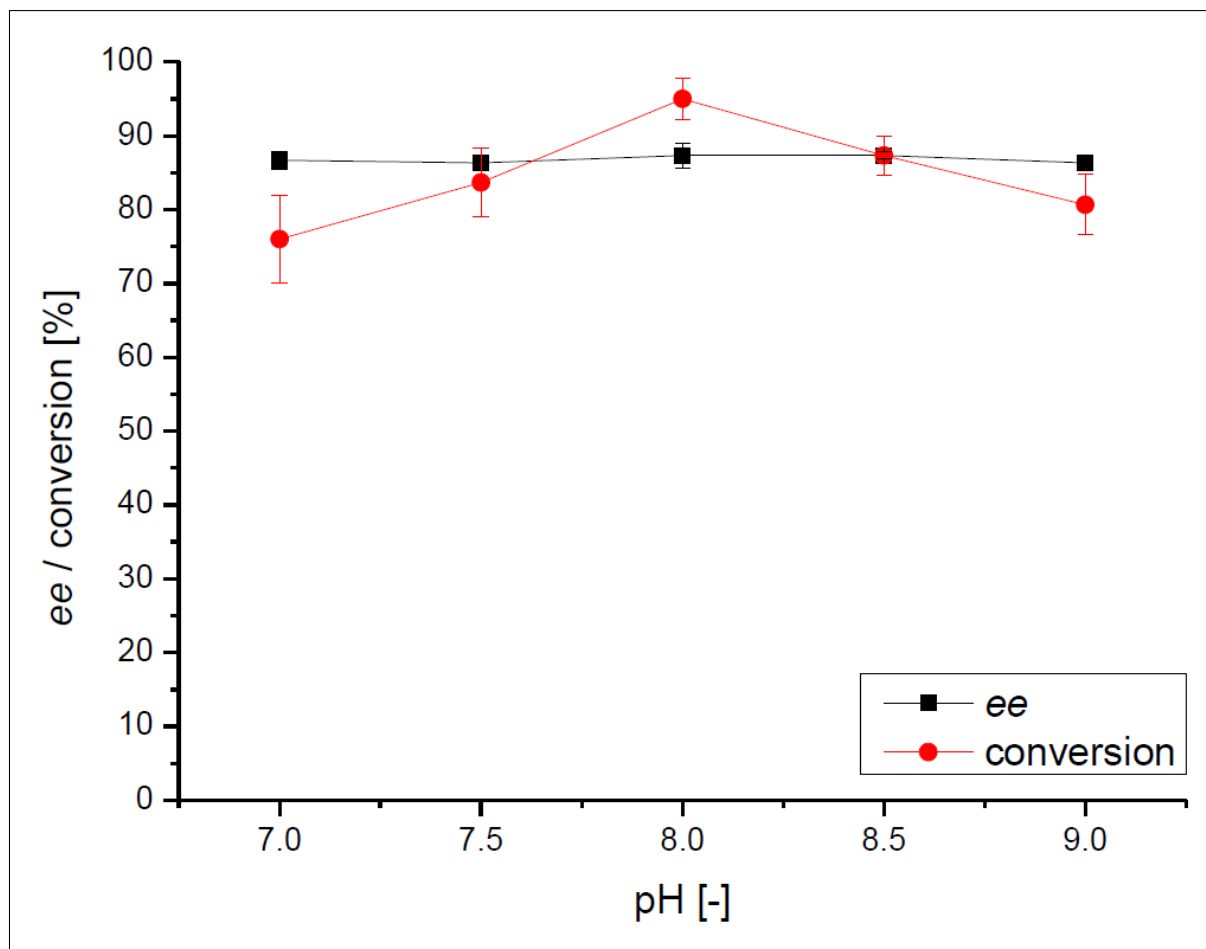


Figure A.4: Influence of pH on the stereoselectivity and conversion of *ApPDC-E469G/T384G/I468A* catalyzed (*S*)-benzoin formation. Reaction conditions: 50 mM TEA buffer, 2 mM MgSO₄, 0.1 mM ThDP, 1 mg mL⁻¹ enzyme, 18 mM benzaldehyde, 20 °C, 800 rpm, 6 h. Lines indicate visual aids.

Consequently, the substrate screening and the synthesis of (*S*)-benzoins in preparative scale (see Publication, Table 4.5) were performed without addition of cosolvents at pH 8.0 and 15 °C using 18 mM of the respective benzaldehyde derivatives.

A.6.7 Product analysis

The identity of (*S*)-benzoin was verified by ¹H and ¹³C NMR spectroscopy. Spectral data were previously reported elsewhere (Demir, Sesenoglu, Eren, et al. 2002). Analytics of phenylacetylcarbinol and 2-hydroxypropiophenone were performed as earlier described (Gehards et al. 2012; Gocke, Graf, et al. 2009).

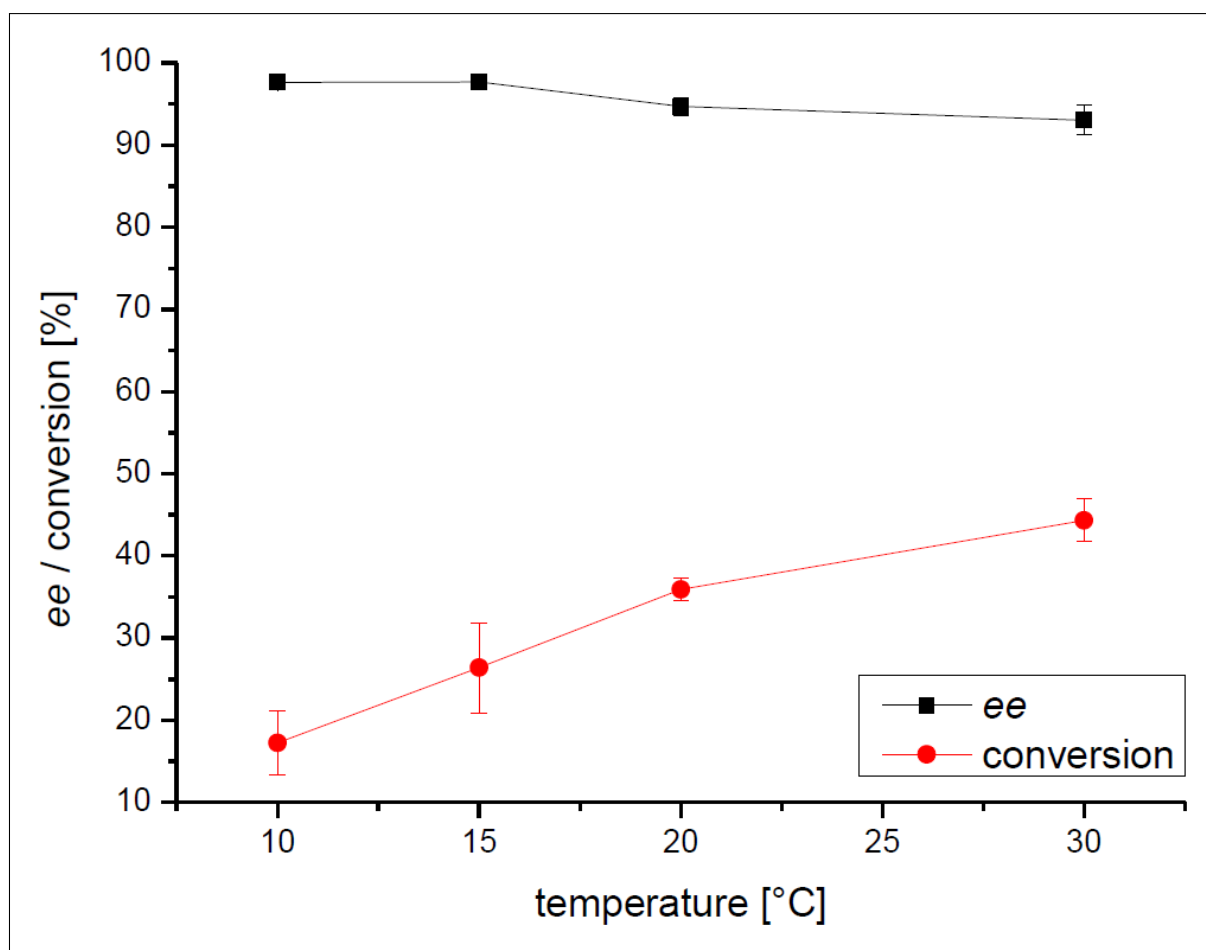


Figure A.5: Influence of temperature on the stereoselectivity and conversion of *ApPDC-E469G/T384G/I468A/W543* catalyzed (*S*)-benzoin formation. Reaction conditions: 50 mM TEA buffer, pH 8.0, 2 mM MgSO₄, 0.1 mM ThDP, 1 mg mL⁻¹ enzyme, 18 mM benzaldehyde, 800 rpm, 6 h. Lines indicate visual aids.

Determination of enantiomeric excess and conversion

The enantiomeric excesses of (*S*)-benzoin (**1a**) and the benzoin derivatives (**1b-1l**) were determined by chiral-phase HPLC using a 1260 Infinity chromatography system (Agilent Technologies, USA) equipped with a Daicel Chiralcel OD-H 5 μm (250 mm×4.6 mm) column (Daicel Chemical IND., France). Products were detected at 254 nm using the eluent *n*-hexane:2-propanol (92:8) at a flow rate of 1.2 mL min⁻¹ and 20 °C. Typical retention times were: **1a**: $t_R(S)$ = 10.1 min, $t_R(R)$ = 14.8 min; **1b**: $t_R(S)$ = 8.5 min, $t_R(R)$ = 11.4 min; **1c**: $t_R(S)$ = 11.5 min, $t_R(R)$ = 14.9 min; **1d**: $t_R(S)$ = 17.5 min, $t_R(R)$ = 25.7 min; **1e**: $t_R(S)$ = 8.5 min, $t_R(R)$ = 11.1 min; **1f**: $t_R(S)$ = 9.0 min, $t_R(R)$ = 12.8 min; **1g**: $t_R(S)$ = 10.3 min, $t_R(R)$ = 14.2 min; **1h**: $t_R(S)$ = 11.5 min, $t_R(R)$ = 15.0 min; **1i**: $t_R(S)$ = 15.7 min, $t_R(R)$ = 23.1 min; **1j**: $t_R(S)$ = 8.8 min, $t_R(R)$ = 9.5 min; **1k**: $t_R(S)$ = 6.5 min, $t_R(R)$ = 9.9 min; **1l**: $t_R(S)$ = 9.7 min, $t_R(R)$ = 22.5 min. Conversions were determined by chiral-phase HPLC (see above) based on the consumption of benzaldehyde or benzaldehyde

derivatives. Typical retention times of benzaldehyde and derivatives were 4.5-5.5 min.

Verification of absolute configuration

The retention times of the respective enantiomers of **1a** could be unambiguously assigned on the basis of commercially available (*S*)-**1a** and (*R*)-**1a**. Enantiomers of **1b-1l** were assigned using the *R*-selective carboligation product catalyzed by *Pf*BAL as a reference (Demir, Sesenoglu, Eren, et al. 2002). In addition, the absolute configurations of **1a** and **1e-1l** were verified by determination of its optical rotations using a Jasco P-2000 polarimeter (Jasco, Germany). **1a**: $[\alpha]_{\text{D}}^{20}$: +154.1 (*c* 1.5, CH₃CN), [Literature value: $[\alpha]_{\text{D}}^{20}$: +138.4 (*c* 0.25, CHCl₃) for 89% ee] (Krawczyk et al. 2004). **1e**: $[\alpha]_{\text{D}}^{20}$: +71.5 (*c* 1.5, CH₃CN); **1f**: $[\alpha]_{\text{D}}^{20}$: +63.9 (*c* 1.5, CH₃CN); **1g**: $[\alpha]_{\text{D}}^{20}$: +49.3 (*c* 1.5, CH₃CN); **1h**: $[\alpha]_{\text{D}}^{20}$: +5.2 (*c* 1.5, CH₃CN); **1i**: $[\alpha]_{\text{D}}^{20}$: +116.4 (*c* 0.25, CH₃CN).


```

gi|504270371 NAMGSQERQHILMVGDSFQLTAQEMAQMVRYKLPVIIIFLVNNRGYVIEIAIHDGPYNYI
gi|489726147 NAMGSQERQHILMVGDSFQLTAQEMAQMVRYKLPVIIIFLVNNRGYVIEIAIHDGPYNYI
gi|517917915 NAMGSQERQHILMVGDSFQLTAQEMAQMVRYKLPVIIIFLVNNRGYVIEIAIHDGPYNYI
gi|529247081 NAMGSQERQHILMVGDSFQLTAQEMAQMVRYKLPVIIIFLVNNRGYVIEIAIHDGPYNYI
gi|489720590 NAMGSQERQHILMVGDSFQLTAQEMAQMVRYKLPVIIIFLVNNRGYVIEIAIHDGPYNYI
gi|20385191 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
gi|178847311 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
gi|178847310 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
gi|178847309 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
gi|178847308 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
gi|178847307 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
gi|178847306 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
gi|178847305 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
gi|178847304 NAMGSQDRQHVVMMVGDGSFQLTAQEVAQMVRYELPVIIIFLINNRGYVIEIAIHDGPYNYI
*****.*.*.*.*:*****.*.*.*.*:*****.*.*.*.*:*****.*.*.*.*:*****.*.*.*.*

```

```

gi|504270371 KNWDYAGLMEVFNAEDGHGLGLKATTAGELEEAIKKAKTNREGPTIIIECQIERSDCTKTL
gi|489726147 KNWDYAGLMEVFNAEDGHGLGLKATTAGELEEAIKKAKANREGPTIIIECQIERSDCTKTL
gi|517917915 KNWDYAGLMEVFNAEDGHGLGLKATTAGELEEAIKKAKANREGPTIIIECQIERSDCTKTL
gi|529247081 KNWDYAGLMEVFNAEDGHGLGLKATTAGELEEAIKKAKANREGPTIIIECQIERSDCTKTL
gi|489720590 KNWDYAGLMEVFNAEDGHGLGLKATTAGELEEAIKKAKANREGPTIIIECQIERSDCTKTL
gi|20385191 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
gi|178847311 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
gi|178847310 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
gi|178847309 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
gi|178847308 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
gi|178847307 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
gi|178847306 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
gi|178847305 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
gi|178847304 KNWDYAGLMEVFNAGEGHGLGLKATTPKELTEAIARAKANTRGPTLIECQIDRTDCTDML
*****.*.*.*.*:*****.*.*.*.* ** ***:***:* ***:***:*.*.*.*.*

```

```

gi|504270371 VEWGKKVAAANSRKPQVS-----
gi|489726147 VEWGKKVAAANSRKPQVS-----
gi|517917915 VEWGKKVAAANSRKPQVS-----
gi|529247081 VEWGKKVAAANSRKPQVS-----
gi|489720590 VEWGKKVAAANSRKPQVS-----
gi|20385191 VQWGRKVASTNARKTTLA-----
gi|178847311 VQWGRKVASTNARKTTLAL EHHHHHHH
gi|178847310 VQWGRKVASTNARKTTLAL EHHHHHHH
gi|178847309 VQWGRKVASTNARKTTLAL EHHHHHHH
gi|178847308 VQWGRKVASTNARKTTLAL EHHHHHHH
gi|178847307 VQWGRKVASTNARKTTLAL EHHHHHHH
gi|178847306 VQWGRKVASTNARKTTLAL EHHHHHHH
gi|178847305 VQWGRKVASTNARKTTLAL EHHHHHHH
gi|178847304 VQWGRKVASTNARKTTLAL EHHHHHHH
*.*.*.*.*:***:*.*.*.* **

```

Table A.9: Family classification in the TEED. Nine superfamilies were established and further subdivided in homologous families.

Superfamily	# homologous families	# proteins	# sequences	# structures
Decarboxylases	57	14828	23067	121
Transketolases	23	10098	14878	40
1-Deoxy-D-xylulose-5-phosphate synthases	3	3743	5801	2
Oxidoreductases	31	8810	12180	11
α -Ketoacid dehydrogenases 1	3	1734	2869	11
α -Ketoacid dehydrogenases 2	34	9818	13376	85
Sulfoxyruvate decarboxylases	6	303	340	0
Phosphonopyruvate decarboxylases	7	384	462	0
α -Ketoglutarate dehydrogenases	4	2847	4520	14

B List of Publications

Scientific articles

1. Vogel, C; Widmann, M; Pohl, M; Pleiss, J. (2012) A standard numbering scheme for thiamine diphosphate-dependent decarboxylases. *BMC Biochemistry* **13**:24-24 [Section 4.1 on pages 76ff., doi:[10.1186/1471-2091-13-24](https://doi.org/10.1186/1471-2091-13-24)]
2. Hailes, HC; Rother, D; Müller, M; Westphal, R; Ward, JM; Pleiss, J; Vogel, C; Pohl, M. (2013) Engineering stereoselectivity of ThDP-dependent enzymes. *FEBS J* **280**:6374-6394 [doi:[10.1111/febs.12496](https://doi.org/10.1111/febs.12496)]¹
3. Westphal, R; Hahn, D; Mackfeld, U; Waltzer, S; Beigi, M; Widmann, M; Vogel, C; Pleiss, J; Müller, M; Pohl, M. (2013) Tailoring (*S*)-selectivity of MenD from *Escherichia coli*. *ChemCatChem* **13**:3587-3594 [doi:[10.1002/cctc.201300318](https://doi.org/10.1002/cctc.201300318)]
4. Gricman, Ł; Vogel, C; Pleiss, J. (2014) Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance. *Proteins* **82**:491-504 [doi:[10.1002/prot.24415](https://doi.org/10.1002/prot.24415)]
5. Westphal, R; Jansen, S; Vogel, C; Pleiss, J; Müller, M; Rother, D; Pohl, M. (2014) MenD from *Bacillus subtilis*: A potent catalyst for the enantiocomplementary asymmetric synthesis of functionalized α -hydroxy ketones. *ChemCatChem* **6**:1082-1088 [doi:[10.1002/cctc.201300690](https://doi.org/10.1002/cctc.201300690)]
6. Vogel, C; Pleiss, J. (2014) The modular structure of ThDP-dependent enzymes. *Proteins* **82** (10):2523-2537 [Section 4.2 on pages 92ff., doi:[10.1002/prot.24615](https://doi.org/10.1002/prot.24615)]
7. Westphal, R; Vogel, C; Schmitz, C; Pleiss, J; Müller, M; Pohl, M; Rother, D. (2014) A Tailor-made chimeric thiamine diphosphate-dependent enzyme for the direct asymmetric

¹This review is not part of the results and discussion presented in this thesis.

synthesis of (*S*)-benzoins. *Angewandte Intl. Ed.* **53** (35):9376-9379 [Section 4.3 on pages 119 ff., doi:[10.1002/anie.201405069](https://doi.org/10.1002/anie.201405069)]

8. Vogel, C; Reusch, W; Pohl, M; Rother, D; Pleiss, J. BioCatNet: a system for the analysis of sequence-structure-function relationships of protein families. *Manuscript in preparation*. [Section 4.4 on pages 127ff.]
9. Gricman, Ł; Vogel, C; Pleiss, J. Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining. *Manuscript in preparation*.

Poster presentations

- Vogel, C; Widmann, M; Pleiss, J. (2013) Functionally relevant residues in ThDP-dependent enzymes. *11th Biotrans*, Manchester (England)
- Vogel, C; Reusch, W; Eberlein, M; Pleiss, J. (2014) Systematic analysis and modeling of functionally relevant residues in ThDP-dependent enzymes. *The 8th International Conference on Thiamine: From Catalysis to Pathology*, Liège (Belgium)
[Best poster award]

Oral presentation

- Vogel, C; Pleiss, J. (2014) Systematic analysis of ThDP-dependent enzymes. *The 8th International Conference on Thiamine: From Catalysis to Pathology*, Liège (Belgium)

Student supervision

Parts of this work were subdivided into smaller projects, which were performed by students.

- Chantal Göttler (Bachelor thesis, 2013), 'Identification of structurally equivalent residues in ThDP-dependent enzymes'
- Lenz Lorenz (Bachelor thesis, 2013), 'The modular structure of aldolases and transaldolases'

-
- Sebastian Enderle (Research internship, 2013), 'Identification of PYR and PP domains in 1-Deoxy-D-xylulose-5-phosphate synthases'
 - Yahayda Aladzeme (Research internship, 2013), 'Identification of PYR and PP domains in α -ketoacid dehydrogenases'
 - Waldemar Reusch (Diploma thesis, 2014), 'Development of a web accessible graphical user interface for the BioCatNet system'
 - Hannah Dienhart (Bachelor thesis, 2014), 'DBUpdate - a tool to update family specific protein databases implemented in the BioCatNet system'

C Erklärung der eigenständigen Arbeit

Hiermit versichere ich, die vorliegende Arbeit selbstständig und ohne Verwendung anderer als der angegebenen Hilfsmittel und Literatur angefertigt zu haben. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Constantin Vogel

Stuttgart, November 2014

The Thiamine Diphosphate (ThDP)-dependent Enzymes form a vast and diverse family of biocatalysts. Due to their potential to catalyze the chemo- and enantioselective formation and cleavage of carbon-carbon bonds, representatives of this enzyme family were well characterized and applied in industrial biotechnology.

However, compared to the tremendous number of amino acid sequences known for this family, structural elucidation and functional characterization was so far limited to a small number of ThDP-dependent enzymes. The main objectives of this work were (1) to provide systems and tools enabling automated, systematic analyses of the large data set on amino acid sequences and (2) to apply those methods in order to investigate the relationships between sequences, structures and biochemical functions of those enzymes.

Within this thesis, a system for the generation and analysis of family-specific protein databases as well as bioinformatical methods were developed. Subsequent application of these tools supported the successful design of ThDP-dependent biocatalysts with the capability to enantioselectively form products that were so far not accessible via enzymatic carbonylation.



University of Stuttgart
Germany

