Data Integration and Data Mining for the Exploration of Enzymatic Sequence-Structure-Function Relationships

Datenintegration und Data-Mining für die Untersuchung enzymatischer Sequenz-Struktur-Funktionsbeziehungen

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

Patrick Christopher Frank Buchholz
aus Waiblingen

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

Tag der mündlichen Prüfung: 7. August 2018

Institut für Biochemie und Technische Biochemie der Universität Stuttgart

2018
Natura non facit saltus.

attributed to Carl von Linné, *Philosophia Botanica* (Stockholm & Amsterdam, 1751)
Acknowledgments

I would like to thank my predecessors for their pioneering work on our relational data model, especially Dr. Markus Fischer, Dr. Constantin Vogel and Waldemar Reusch for their previous versions of the database system and their kind assistance. Experimental data and test cases for the thiamine diphosphate-dependent enzymes engineering database (TEED) were provided generously by Dr. Anna Baierl, Saskia Bock, Stephan Malzacher, Rüdiger Ohs and Matthias Schapfl, as starting points for the currently ongoing research project that is aimed at extending the BioCatNet database system. Said research project is conducted in collaboration with Prof. Dr. Selin Kara and Jennifer Engel whose suggestions on usability and data exchange are highly appreciated.

I have to thank Rüdiger Ohs for providing the computational tools for kinetic parameter estimations used herein, as well as his supervisor Prof. Dr.-Ing. Antje Spieß, for the detailed discussions on the applications and interpretations of kinetic modeling. I also acknowledge the further members of the research group FOR1296, especially Prof. Dr. Martina Pohl and Prof. Dr. Dörte Rother, for their suggestions improving the usability of the BioCatNet database system.

I thank Dr. Silvia Fademrecht for her comprehensive collections of pre-processed sequence data sets. I thank Oliver Buß, Maike Gräff and Peter Klausmann for their effort in collecting and sorting the large data on protein sequences, structures and literature information for ω-transaminases leading us to a new family-specific protein database, the ω-transaminase engineering database (oTAED). The suggestions of Dr. Jens Rudat on ω-transaminases are appreciated.

I thank Christine Klump-Klug for her patient support in everyday organization tasks at the institute. Lenz Lorenz, Simone Prömel and Catharina Zeil implemented and applied helpful bioinformatic tools for the analysis of sequence data. Furthermore, I thank our former student assistants Jannik Seidel and Marco Zimmer for their tests on the analysis of protein sequence networks. I thank Prof. Dr. Uta Freiberg for the inspiring discussions on the mathematical interpretation of protein sequence networks.

I would like to thank Prof. Dr. Bernhard Hauer for the opportunity to work at the institute and for taking over the chair of my examination, as well as Prof. Dr.-Ing. Ralf Takors for his contribution as examiner. I sincerely thank Prof. Dr. Jürgen Pleiss for the opportunity to participate in the research projects that contributed to this work, the fruitful and inspiring discussions on the respective results and his helpful suggestions on the manuscripts.

Finally, computational work depends on reliable hardware and infrastructure that could be maintained and extended thanks to many helping hands from the whole Bioinformatics group. I would like to sincerely thank my attentive colleagues Catharina Zeil and Mahyar Ghermezi.
for their support in the administration of our demanding database infrastructure.

I cordially thank my parents, Frank and Regina, and my grandfather Richard for their loving support and assistance in all phases of this work.
Contents

Acknowledgments 3

Contents 5

List of Figures 8

List of Tables 16

Abbreviations 19

Parameters and variables 21

Abstract 23

Zusammenfassung 25

1 Introduction 28

1.1 Data on sequence, structure and function of enzymes .......................... 28

1.2 Sequence-sequence relationships: networks ........................................... 30

1.2.1 Exemplary properties of networks ................................................. 31

1.3 Sequence-structure relationships: standard numbering schemes .......... 32

1.4 Exemplary enzymes and enzyme families ............................................. 33

1.4.1 \( \omega \)-transaminases ................................................................. 33

1.4.2 Benzaldehyde lyase: a ThDP-dependent decarboxylase ..................... 34

1.5 Kinetic model equations for symmetric carboligation .......................... 35

1.6 Data integration and data mining ....................................................... 38

2 Results 40

2.1 Properties of protein sequence networks ............................................. 40

2.1.1 Distributions of degrees and network dimensions .............................. 41

2.1.2 Distributions of community sizes .................................................. 46

2.1.3 Exemplary sequence networks of \( \omega \)-transaminases ........................ 49

2.2 Kinetic modeling in the BioCatNet database system ............................. 51

2.2.1 Extensions of the relational data model ......................................... 51

2.2.2 Parameter estimations for \( P/f \)BAL-catalyzed symmetric carboligation .. 53
3 Discussion

3.1 From protein families to protein networks

3.1.1 Connectedness of protein sequence space

3.1.2 Hub regions in protein sequence space

3.1.3 Evolutionary constraints for protein sequences

3.2 From data management to model management

3.2.1 Models for bioinformatic sequence analyses

3.2.2 Managing time-course data and kinetic models

3.2.3 Data integration for bioinformatics and white biotechnology

4 List of publications and manuscripts

5 Bibliography

6 Erklärung der eigenständigen Arbeit

A Appendix: Publications

A.1 The scale-free nature of protein sequence space

A.1.1 Abstract

A.1.2 Introduction

A.1.3 Methods

A.1.4 Results

A.1.5 Discussion

A.1.6 Supporting Information

A.2 Percolation in protein sequence space

A.2.1 Abstract

A.2.2 Introduction

A.2.3 Materials and methods

A.2.4 Results

A.2.5 Discussion

A.2.6 Supporting Information

A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

A.3.1 Abstract

A.3.2 Introduction

A.3.3 Methods

A.3.4 Results

A.3.5 Discussion
A.3.6 Supporting Information ................................................. 170
A.4 BioCatNet: a database system for the integration of enzyme sequences and bio-
catalytic experiments ......................................................... 174
A.4.1 Abstract ................................................................. 174
A.4.2 Introduction ............................................................. 174
A.4.3 Results and Discussion .................................................. 177
A.4.4 Conclusion ............................................................... 181
A.4.5 Computational Methods ................................................... 181
A.4.6 Supporting Information ............................................... 187
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-
catalyzed self-ligation .......................................................... 194
A.5.1 Abstract ................................................................. 194
A.5.2 Introduction ............................................................. 194
A.5.3 Experimental section ..................................................... 196
A.5.4 Results ................................................................. 200
A.5.5 Discussion .............................................................. 202
A.5.6 Supporting Information ............................................... 210
List of Figures

1.1 Symmetric carboligations of benzaldehyde to benzoin (upper panel) and 3,5-dimethoxybenzaldehyde to 3,3′,5,5′-tetramethoxybenzoin (lower panel) catalyzed by benzaldehyde lyase from *Pseudomonas fluorescens*. .......................... 35

1.2 Representations inspired by the systematic of Cleland (1963) for the Michaelis-Menten model for symmetric carboligation (model 1) with substrate (A), product (P), free enzyme (E) and enzyme-substrate/product-complexes (EA, EAA, EP). For the models of the ordered bi-uni mechanism (models 2 to 4), it is assumed that for the first and second binding substrate (donor A and acceptor B) $k_1 = k_2$ and $k_{-1} = k_{-2}$ (Zavrel et al., 2008; Ohs et al., 2018). .......................... 36

2.1 Distribution of the number of sequences $N$ in a distance-based network of TEM β-lactamases having $n$ first neighbors. The degree distribution follows a power law with exponent $\gamma = 1.2$. Compare with Figures A.6 and A.7 (see also Figures A.1 and A.2). ................................................................. 41

2.2 Degree distributions for the protein families with low microdiversity from Table 2.1 with neighbors defined by $\geq 95\%$ global sequence identity. The corresponding scale-free exponents $\gamma$ were derived from linear regression for degrees $\leq 50$ (bHAD, DC) or $\leq 70$ (oTA, SDR) and are summarized in Table 2.1. See also Figure A.4. ................................................................. 43

2.3 Cumulative distributions of sequence pairs $p(d)$ for pairwise distances of $d\%$ for the protein families TEM (open squares), DC (filled squares), bHAD (open circles), SDR (diamonds) and oTA (filled circles) from Table 2.1 in subsequent distance intervals of 2 % distance $d$ (100 % - sequence identity in %). Linear fits are shown as red lines for distances up to 10 % identity (up to 4 % for TEM). For further distances between 30 and 70 %, an approximately linear area is depicted in red (compare with Figure 2.4). See also Figure A.5. ................................................................. 44

2.4 Detail view from Figure 2.3 for the protein families DC (filled squares), bHAD (open circles), SDR (diamonds) and oTA (filled circles) showing cumulative distributions of sequence pairs $p(d)$ for pairwise distances of $d\%$. Regression lines are given for the sequence pairs from the DC (blue line), bHAD (red line), SDR (blue dashed line) and oTA family (red dashed line). ................................................................. 45

2.5 Distribution of community sizes, resulting from subgroups clustered by 60 % pairwise sequence identity determined by USEARCH (Edgar, 2010), for the protein families SDR, oTA, DC and bHAD following a power law distribution according to Equation 2.1. Compare with additional results in Figure A.14. .......... 46
2.6 Fisher exponent $\tau$ of the community size distributions of protein superfamilies from Figure 2.5 for clustering thresholds between 60 and 90% pairwise sequence identity with extrapolated Fisher exponent $\tau_{100}$ determined by linear regression. Compare with additional results in Figure A.15.

2.7 Example of a protein sequence network for 447 representative sequences from the superfamily of $\omega$-transaminases from fold type IV, clustered by 30% sequence identity in USEARCH (Edgar, 2010) with 13,752 edges defined by a threshold of 50% sequence similarity (derived from pairwise alignments according to Needleman and Wunsch (1970). Sequences are annotated as 4-amino-4-deoxychorismate lyase (blue), (R)-amine transaminase (green), L-branched chain amino acid aminotransferase (red) or D-amino acid aminotransferases (black). See also Figure A.24.

2.8 Detail view of the BioCatNet relational data model focusing on tables for experimentally determined ratios (like yield, conversion or enantiomeric excess ($ee$)).

2.9 Fits for the dimethoxy-benzaldehyde self-ligation and the microkinetic model with substrate-dependent enzyme inactivation (model 4, lines). Symbols as described in Figure A.36. The respective parameter estimates are given in Table 2.3 with reaction conditions according to Figure A.36 and Table A.15 (Equivalent with Figure A.44).

2.10 Fits for the benzaldehyde self-ligation and the microkinetic model with substrate-dependent inactivation (model 4, lines). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15 (Equivalent with Figure A.45).

3.1 Concept for data and model integration in a common data- and modelbase on the levels of protein sequence (1), structure (2) and experimental time-course data (3). Models refer to concepts and assumptions that are applied on the original data, such as: alignments that are applied on protein sequences to construct protein sequence networks, standard numbering schemes (based on profile HMMs of structure-guided sequence alignments) or equations of kinetic models. The application of models results in model-dependent information (compare with Figure A.28) that cannot be reproduced and interpreted without annotations on the original data and the models themselves. Thus, routines of data mining rely on both data and model quality. Exemplary results of data mining could be the identification of hubs or bridges in protein sequence networks, conservation analyses based on standard numbering schemes or kinetic parameter estimations.
A.1 Sequence network for 267 TEM β-lactamases formed by 401 point mutations (edges) with 259 sequences forming a densely connected network with two hub sequences (TEM-1 depicted as black rectangle, TEM-116 as black oval). First neighbors of hub sequences are depicted in dark gray, other sequences in white.

A.2 Distribution of the number of sequences \( N \) in a network of TEM β-lactamase point mutations having \( n \) first neighbors (Figure A.1). The degree distribution follows a power law with exponent \( \gamma = 1.2 \).

A.3 Distributions of pairwise global sequence identity for the protein families from Table A.1 as determined by high-scoring sequence pairs in USEARCH (Edgar, 2010).

A.4 Neighbor distribution for the protein families with low microdiversity from Table A.1 with neighbors defined by \( \geq 95\% \) global sequence identity. The corresponding scale-free exponents \( \gamma \) were derived from linear regression for degrees \( \leq 50 \) (bHAD, DC) or \( \leq 70 \) (oTA, SDR) and are summarized in Table A.2.

A.5 Cumulated distributions of sequence pairs \( p(d) \) for pairwise distances of \( d\% \) of the protein families TEM (open squares), DC (filled squares), bHAD (open circles), SDR (diamonds) and oTA (filled circles) from Table A.1 in subsequent distance intervals of 2\% distance \( d \) (100\% - sequence identity). Linear fits are shown as red lines for distances up to 10\% identity (up to 5\% for TEM). For further distances between 70 and 30\%, an approximately linear area is depicted in red.

A.6 Sequence network for 267 TEM β-lactamases connected by 401 edges above a 99.5\% pairwise sequence identity threshold, in comparison to the point mutation network (Figure A.1). Hub sequences are depicted in black (TEM-1 as black rectangle, TEM-116 as black oval) with their first neighbors depicted in dark gray, other sequences in white.

A.7 Distribution of the number of sequences \( N \) having \( n \) first neighbors for the distance-based network of TEM β-lactamases (Figure A.6). The degree distribution hints at a power law distribution with exponent \( \gamma = 1.2 \) (A). In addition, probability density functions were fitted for a power-law distribution (line, \( \gamma = 1.2 \)), a Gaussian distribution (dashed line, \( \mu = 3.0, \sigma = 6.4 \)) and a Poisson distribution (dotted line, \( \lambda = 3.0 \)) with residual sum of squares 0.01, 0.2 and 0.1, respectively (B - D).

A.8 Degree distribution for the protein families with low microdiversity from Table A.1 with neighbors defined by \( \geq 90\% \) global sequence identity. Linear regression was performed for degrees \( \leq 50 \) (bHAD, DC) or \( \leq 70 \) (oTA, SDR).
A.9 Degree distribution for the protein families with low microdiversity from Table A.1 with neighbors defined by \( \geq 85\% \) global sequence identity. Linear regression was performed for degrees \( \leq 50 \) (bHAD, DC) or \( \leq 70 \) (oTA, SDR).

A.10 Degree distribution for the protein families with low microdiversity from Table A.1 with neighbors defined by \( \geq 80\% \) global sequence identity. Linear regression was performed for degrees \( \leq 50 \) (bHAD, DC) or \( \leq 70 \) (oTA, SDR).

A.11 Cumulative distributions of sequence pairs \( p(d) \) for pairwise distances of \( d \% \) of the DC protein superfamily for the complete data set (open squares) and 10\%, 20\%, \ldots, 90\% randomly selected subsets of sequences (filled squares). The areas marked in red correspond to the linear approximations from Figure A.5.

A.12 Cumulative distribution of sequence pairs \( p(d) \) with distance in \( d \) point mutations of TEM \( \beta \)-lactamases (open squares). Linear fits are shown for \( d = 1, 2 \) (red line) and for \( d = 2, 3, 4 \) point mutations (blue line).

A.13 Distributions of pairwise global sequence identity for the protein families of \( \alpha/\beta \) hydrolases (abH), short-chain dehydrogenases/reductases (SDR), \( \omega \)-transaminases (oTA), cytochrome P450 monoxygenases (CYP), thiamine diphosphate-dependent decarboxylases (DC) and \( \beta \)-hydroxyacid dehydrogenases/imine reductases (bHAD).

A.14 Cluster size distribution of \( \alpha/\beta \) hydrolases (abH), short-chain dehydrogenases/reductases (SDR), \( \omega \)-transaminases (oTA), cytochrome P450 monoxygenases (CYP), thiamine diphosphate-dependent decarboxylases (DC), and \( \beta \)-hydroxyacid dehydrogenases/imine reductases (bHAD) follow a power law distribution: \( N(s) \sim s^{-\tau} \) (\( N(s) \), number of clusters of size \( s \); \( \gamma \), Fisher exponent). Cluster criterion: 60\% global sequence identity.

A.15 Fisher exponent \( \tau \) of the size distribution of homologous families for clustering cutoffs between 60 and 90\% with extrapolated Fisher exponent \( \tau_{100} \) determined by linear regression (abbreviations according to Figure A.14).

A.16 Model distributions displayed as log-log plot: Gaussian distribution \( N(s) = a \cdot \exp\left(-\frac{1}{2}\frac{(s-\mu)^2}{\sigma^2}\right) \) with \( a = 10000, \mu = 200, \sigma = 50 \), exponential distribution \( N(s) = a \cdot \exp(-bs) \) with \( a = 10000 \) and \( b = 0.2 \), power law distribution \( N(s) = a \cdot s^{-\tau} \) with \( a = 10000, -\tau = 2.5 \).

A.17 Cluster size distributions for 60, 70, 80, and 90\% global sequence identity of the six protein superfamilies from Table A.6 (\( \alpha/\beta \)-hydrolases in blue, shortchain dehydrogenases/reductases in green, \( \omega \)-transaminases in black, cytochrome P450 monoxygenases in red, thiamine diphosphate-dependent decarboxylases in cyan and \( \beta \)-hydroxyacid dehydrogenases/imine reductases in orange).
A.18 Cluster size distributions for 60, 70, 80, and 90% global sequence identity of all abH sequences (filled squares) and randomly selected abH sequences: 50% (open squares), 25% (filled circles) and 12.5% (open circles) of the original dataset.

A.19 Details of sequence identity networks for two homologous families of short-chain dehydrogenases/reductases (SDR) with clustering cutoff at 39% sequence identity. The network shows bridges connecting the two homologous families (indicated in red hexagons). Visualization in Cytoscape (version 3.2.1) using organic layout.

A.20 Details of sequence identity networks for two homologous families of short-chain dehydrogenases/reductases (SDR) with clustering cutoff at 40% sequence identity. The bridge sequences from Figure A.19 are indicated in red hexagons. Visualization in Cytoscape (version 3.2.1) using organic layout.

A.21 Salt bridge in ω-TA Fold type I. Shown for 4AOA, 2YKU, 5GFH, and 3NUI. The conserved D210 seems to be an important salt bridge starting point, but the corresponding salt bridge partner residue is not conserved, the distance between arginine/asparagine and aspartate is < 3.5 Å.

A.22 The active sites of (R)- and (S)-selective ω-TAs (Fold type IV and I, respectively) as viewed from the re- and si-face, respectively. The functional residues were defined according to Lyskowski et al. (2014) and to Humble et al. (2012). For (S)-selective-ω-TA, the amino acid residues of the phosphate binding cup at position 82 are serine or threonine (Rice et al., 2000; Denesyuk et al., 2003).

A.23 Structure comparison of Fold type I and IV. The pattern of Fold type I consists of an α/β/α pattern with the active site located at the interface of the homodimer (only monomer is shown). Fold type IV consists of 2 clear separated domains. Domain 1 is a 2 layer sandwich. Domain 2 consists of an α-β barrel. The active site is also located at the interface of the homodimer.

A.24 Protein sequence networks of Fold type IV sequences with (A) annotated sequences matching positions of likely (R)-selective transaminases depicted in black and with (B) motifs for ADCL (blue), DATA (black), l-BCAT (red), and ATA (green) from 27 and Table A.13. Nodes correspond to representative sequences of clusters formed by 30% identity in USEARCH. A cutoff of 50% pairwise sequence similarity is used to select the edges. The network is shown as force-directed layout, with pairs with higher similarity arranged in closer proximity.

A.25 Distribution of sequence lengths of Fold type IV (oTA IV) and Fold type I (oTA I), ω-transaminases from the oTAED, represented as histograms and boxplots with maximum whisker length 1.5 times the interquartile range and outliers marked as crosses.
A.26 ω-transaminases of Fold type I (A) and Fold type IV (B) from psychrophilic (green), thermophilic (red), and hyperthermophilic (black) sources retrieved from the BacDive database (release 27.02.2017) (Söhngen et al., 2014) and the representative node (orange) containing the motif from Stekhanova et al. (2017) for thermostable Fold type IV sequences (V/I)xLDxR and PFG(K/H)YL. The nodes correspond to representative sequences of clusters formed by 30% identity in USEARCH (Edgar, 2010). A cutoff of 50% pairwise sequence similarity was used to construct the edges. The network is shown as force-directed layout, thus pairs with higher similarity are positioned in closer proximity.

A.27 Workflow between bioinformatics and experiment using the BioCatNet database system. Arrows pointing from left to right represent data analysis or modeling, arrows pointing from right to left represent submission of biocatalytic data via file upload. Dashed arrows mark initial procedures and modeling steps beyond the scope of the database.

A.28 Concept of the established BioCatNet database system. Original data are stored in the individual enzyme family database on the levels of protein sequence, structure and experimental data. Analysis of data leads to new knowledge which helps to amend the database (model-derived information). Sequences and structures are parsed from other repositories whereas experimenters contribute experimental data and sequences of their applied enzymes. Expansions of the workflow are currently under development to include metagenomic sequences and establish a framework for estimation of kinetic parameters (*).

A.29 Screenshot detail from the BioCatNet web interface depicting a C-terminally His-tagged variant of pyruvate decarboxylase from Acetobacter pasteurianus. The Mutated residue (G469), domain boundaries (PYR: pyrimidine interaction, PP: pyrophosphate interaction) and the His-tag are annotated. “DC” indicates the respective standard position number (Vogel et al., 2012).

A.30 Exemplary screenshot detail from the BioCatNet web interface, showing the possibility to display a plain text summarizing the reaction conditions from the individual tables.

A.31 Screenshot detail from the BioCatNet web interface for a measurement of substrate and product concentrations over time (here: self-ligation from benzaldehyde to benzoin).

A.32 Simplified scheme of the relational data model for the BioCatNet database system for experimental data. The PARAMETER_MEASUREMENTS and ENZYME_FEEDS tables link to SEQUENCES, which was inherited from the previously described DWARF system (Fischer et al., 2006).
A.33 Alignment of 30 entries in the NCBI protein database for “pyruvate decarboxylase from Acetobacter pasteurianus” showing multiple sequences varying both at the N- and C-termini and within the sequences (beginning on page 190). The alignment was generated using Clustal omega (Sievers et al., 2011).

A.34 Simplified scheme of the workflow that combines experimental data retrieved from the BioCatNet database with routines for the estimation of kinetic parameters. In each repetition of the kinetic parameter estimation, three subsequent steps of parameter estimation are performed as a trade-off between accuracy and performance (Ohs et al., 2018), using previously interpolated reaction rates (Monte-Carlo and algebraic parameter estimation) and finally using progress curves of concentrations over time (dynamic parameter estimation). Quality indicators comprise standard deviations of kinetic parameters and the residual sum of squares (RSS) with respect to the full data set.

A.35 Microkinetic representation of the Michaelis-Menten model for symmetric carboligation (model 1) with substrate (A), product (P), free enzyme (E), and enzyme-substrate/product-complexes (EA, EAA, EP). For the microkinetic model of the ordered bi-uni mechanism (models 2 to 4), it is assumed that for the first and second binding substrate (A and B) $k_1 = k_2$ and $k_{-1} = k_{-2}$ (Zavrel et al., 2008; Ohs et al., 2018).

A.36 Fits of the two-substrate Michaelis-Menten model (model 1, dotted lines) to the full progress curve of dimethoxy-benzaldehyde self-ligation reactions (red points (Zavrel et al., 2008)). The reaction was performed in 50 mM KPi buffer with 0.25 mM MgCl$_2$, 0.25 mM ThDP and 30% DMSO at pH 8.5 and 25°C. The initial concentrations of substrate and enzyme are given in Table A.15. The respective parameter estimates are given in Table A.18.

A.37 Fits for the benzaldehyde self-ligation (Ohs et al., 2018) and the two-substrate Michaelis-Menten model (model 1). Fitted data are indicated as lines, experimental measurement data as points with straight lines and full circles for the substrate benzaldehyde and dotted lines and open circles for the product benzoin. The respective parameter estimates are given in Table A.18. The reaction was performed in 50 mM TEA buffer with 2.5 mM MgSO$_4$, 0.5 mM ThDP and 30% DMSO at pH 8.5 and 30°C. The initial concentrations of substrate and enzyme are given in Table A.15.

A.38 Fits for the first six minutes of the dimethoxy-benzaldehyde self-ligation (points) and the two-substrate Michaelis-Menten model (model 1, dotted lines). The respective parameter estimates are given in Table A.16 with reaction conditions according to Figure A.36 and Table A.15.
A.39 Fits for the first six minutes of the benzaldehyde self-ligation and the two-substrate Michaelis-Menten model (model 1). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15. 215

A.40 Fits for the first six minutes of the dimethoxy-benzaldehyde self-ligation (points) and the macrokinetic model (model 2, dotted line). The respective parameter estimates are given in Table A.16 with reaction conditions according to Figure A.36 and Table A.15. 216

A.41 Fits for the benzaldehyde self-ligation and the macrokinetic model (model 2). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15. 217

A.42 Fits for the dimethoxy-benzaldehyde self-ligation (points) and the microkinetic model (model 3, dotted line). The respective parameter estimates are given in Table A.16 with reaction conditions according to Figure A.36 and Table A.15. 218

A.43 Fits for the benzaldehyde self-ligation and the microkinetic model (model 3). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15. 219

A.44 Fits for the dimethoxy-benzaldehyde self-ligation (points) and the microkinetic model with substrate-dependent enzyme inactivation (model 4, dotted lines). The respective parameter estimates are given in Table A.16 with reaction conditions according to Figure A.36 and Table A.15. 220

A.45 Fits for the benzaldehyde self-ligation and the microkinetic model with substrate-dependent inactivation (model 4). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15. 221
List of Tables

2.1 Overview of the analyzed protein family networks by number of nodes, edges and maximal degree (number of neighbors) for a 95% sequence identity threshold, with average sequence length. The small family of TEM β-lactamases is shown as reference, due to its high microdiversity, with a threshold of 99.5% sequence identity (a). The enzyme families comprise TEM β-lactamases (TEMs), β-hydroxyacid dehydrogenases/imine reductases (bHADs), thiamine diphosphate-dependent decarboxylases (DCs), ω-transaminases (oTAs) and short-chain dehydrogenases/reductases (SDRs). Values for γ refer to the scaling exponents of the degree distributions from Figures 2.1 and 2.2. Values for $D_f$ refer to the slope in Figure 2.3. Compare with Table A.2. 42

2.2 Protein superfamily size and the Fisher exponent extrapolated to 100% sequence identity ($\tau_{100}$) of the four protein families. Compare with additional results in Table A.6. 47

2.3 Estimated kinetic parameters and the residual sum of squares ($RSS$) over sample size $n$ for the benzaldehyde lyase-catalyzed self-ligation of 3,5-dimethoxybenzaldehyde and four kinetic models. Estimated parameter values are given in bold, whereas parameter values obtained from Equations 1.14 to 1.21 are set in regular font. The relative standard deviations (in %) are given in parentheses. Relative standard deviation values $< 100\%$ are marked with * to indicate the higher reliability of the respective kinetic parameter values. 56

2.4 Estimated kinetic parameters and the residual sum of squares ($RSS$) over sample size $n$ for the benzaldehyde lyase-catalyzed self-ligation of benzaldehyde and four kinetic models. Footnotes and descriptions are equivalent to Table 2.3. 57

A.1 Overview of the analyzed protein family networks by number of nodes (sequences) and maximal degree (number of neighbors) for a 95% sequence identity threshold, with average sequence length. The small family of TEM β-lactamases is shown as reference due to its high microdiversity with a threshold of 99.5% sequence identity (a). 116

A.2 Overview of the analyzed protein families from Table A.1 and their derived parameters. The scale-free exponent $\gamma$ refers to sequence identity networks constructed with pairwise identity thresholds of 95% (compare with Table A.4, 99.5% threshold for TEM β-lactamases$^a$). Network dimension $D_f$ refers to the slope from Figure A.5 in different regions of pairwise sequence identity ($> 90\%$). 116
A.3 Exemplary network hubs and their annotations from sequence networks with a threshold of 95% sequence identity (99.5% for TEM β-lactamases) for the protein families from Table A.1.

A.4 Scaling exponents γ for randomly selected subnetworks of the DC superfamily, with edges formed by a threshold of 95% pairwise sequence identity. Linear regressions were performed up to a limited number of neighbors only, due to low sampling quality for higher degrees. Thus, values for γ were determined up to a maximum degree.

A.5 Exemplary protein sequences found in hub regions of the DC networks for varying subsets of randomly selected sequences. The Annotations are listed as ”pyruvate dihydrogenase subunit” (PDH), ”glyoxylate carboligase” (GLX) or ”acetolactase synthase 2 catalytic subunit” (ALS). Pairwise sequence identities towards the hub sequence of the complete network (WP_044256366) are given in the column on the right.

A.6 Protein superfamily size and the Fisher exponent extrapolated to 100% sequence identity (τ_{100}) of the six protein families.

A.7 Conserved positions in putative Fold type I (S)-selective ω-TA sequences with standard numbering according to the ω-TA from Mesorhizobium sp. LUK (PDB accession 2YKU) and their location inside the protein structure or annotated function. Positions listed here are conserved to at least 70%.

A.8 Conserved positions in putative Fold IV ω-TA sequences with standard numbering according to the ω-TA from Aspergillus terreus (PDB accession 4CE5) annotated as in Table A.7.

A.9 Substrate specificity-determining positions and substrate-specific sequence motifs in Fold type I (S)-selective ω-transaminases. Standard positions refer to position numbers of the ω-TA from Mesorhizobium sp. LUK (PDB accession 2YKU).

A.10 Substrate specificity-determining positions and substrate-specific sequence motifs in Fold type IV (R)-selective ω-transaminases. Standard positions refer to position numbers of the ω-TA from Aspergillus terreus (PDB accession 4CE5).

A.11 Query sequences for the initial BLAST search to set up the ω-Transaminase Engineering Database (oTAED).

A.12 Reference structures for the standard numbering schemes of ω-transaminase superfamilies Fold Type I and Fold Type IV.
A.13 Relevant fingerprints of oTAED Fold Type IV standard positions as reported previously (Höhne et al., 2010) for 4-amino-4-deoxychorismate lyase (ADCL), d-amino acid aminotransferases (DATA), L-branched chain amino acid aminotransferases (L-BCAT) and (R)-amine transaminases (ATA). In addition, the catalytic lysine at Fold Type IV standard position 180 is conserved in the four subfamilies. .................................................. 173

A.14 Overview of mandatory and optional input for submitting experimental data to BioCatNet. Experimental data can be experimenter-defined or measured concentrations as well as derived properties like conversion, yield or enantiomeric excess. .................................................. 186

A.15 Initial experimental conditions. .................................................. 207

A.16 Estimated kinetic parameters and the residual sum of squares (RSS) over sample size $n$ for the benzaldehyde lyase-catalyzed self-ligation of 3,5-dimethoxybenzaldehyde and four kinetic models. Estimated parameter values are given in bold, whereas parameter values obtained from Equations A.10 to A.20 are set in regular font. The relative standard deviations (in %) are given in parentheses. Relative standard deviation values $< 100 \%$ are marked with $^*$ to indicate the higher reliability of the respective kinetic parameter values. .................................................. 208

A.17 Estimated kinetic parameters and the residual sum of squares (RSS) over sample size $n$ for the benzaldehyde lyase-catalyzed self-ligation of benzaldehyde and four kinetic models. Footnotes and descriptions are equivalent to Table A.16. .............................. 209

A.18 Estimated kinetic parameters for two substrates and the Michaelis-Menten model. Relative standard deviations ( %) are given in italics. The residual sum of squares (RSS) over sample size $n$ is indicated for nine experiments containing $n = 2786$ samples for 3,5-dimethoxy-benzaldehyde self-ligation (Zavrel et al., 2008) and thirteen experiments containing $n = 374$ samples for benzaldehyde self-ligation (Ohs et al., 2018). .................................................. 222
Abbreviations

Å Ångström (1 Å = 0.1 nm)

abH α/β-hydrolase

ADCL 4-amino-4-deoxychorismate lyase

ApPDC pyruvate decarboxylase from *Acetobacter pasteurianus*

ATA (R)-amine transaminase

BA benzaldehyde

BAL benzaldehyde lyase

BCAT (L-)branched chain amino acid aminotransferase

bHAD β-hydroxyacid dehydrogenase/imine reductase

BLAST Basic Local Alignment Search Tool

BRENDA Braunschweig Enzyme Database

CYP cytochrome P450 monooxygenase

DATA d-amino acid aminotransferase

DC thiamine diphosphate-dependent decarboxylase

DDBJ DNA Data Bank of Japan

DMBA 3,5-dimethoxy-benzaldehyde

DMSO dimethyl sulfoxide

DWARF data warehouse system for analyzing protein families

ee enantiomeric excess

GABA γ-aminobutyric acid

GO Gene Ontology

Hfam homologous family

HMM hidden Markov model
HPLC  high performance liquid chromatography
IPA  isopropyl amine
IUPAC  International Union of Pure and Applied Chemistry
KEGG  Kyoto Encyclopedia of Genes and Genomes
KPi  potassium phosphate
L  liter
LED  lipase engineering database
LIMS  Laboratory Information Management System
M  mol per liter, \( \frac{\text{mol}}{L} \)
NCBI  National Center for Biotechnology Information
\( \omega \)-TA  \( \omega \)-transaminase
\( \omega \)-TAED  \( \omega \)-transaminase engineering database
PDB  Protein Data Bank
PfBAL  benzaldehyde lyase from \textit{Pseudomonas fluorescens}
PLP  pyridoxal 5’-phosphate
rpm  rotations per minute
RSS  residual sum of squares
SDR  short-chain dehydrogenase/ reductase
SMILES  Simplified Molecular-Input Line-Entry System
SSN  Sequence Similarity Network
STAMP  Structural Alignment of Multiple Proteins
STRENDAs  Standards for Reporting Enzymology Data
TA  transaminase
TEA  triethanolamine
**Parameters and variables**

β critical exponent for percolation (dependent on lattice geometry)

c<sub>A</sub> concentration of donor substrate A

c<sub>B</sub> concentration of acceptor substrate B

c<sub>Et</sub> total enzyme concentration

c<sub>P</sub> product concentration

D<sub>f</sub> approximation for (fractal) network dimension

d pairwise distance between sequences

γ scaling exponent

k<sub>cat</sub> turnover number in \( \frac{1}{min} \)

k<sub>catf</sub> turnover number of the forward reaction in \( \frac{1}{min} \)

k<sub>i</sub> microkinetic parameter, elementary rate constant

k<sub>1</sub> rate constant for first substrate binding in \( \frac{l}{mmol·min} \)

k<sub>−1</sub> rate constant for first substrate release in \( \frac{1}{min} \)

k<sub>2</sub> rate constant for second substrate binding in \( \frac{l}{mmol·min} \)

k<sub>−2</sub> rate constant for second substrate release in \( \frac{1}{min} \)

k<sub>3</sub> rate constant for product release in \( \frac{1}{min} \)

k<sub>−3</sub> rate constant for product binding in \( \frac{l}{mmol·min} \)
\( K_{eq} \) equilibrium constant in \( \frac{L}{mmol} \)

\( K_{iA} \) inhibition constant of donor substrate A in \( \frac{mmol}{l} \)

\( K_{iB} \) inhibition constant of acceptor substrate B in \( \frac{mmol}{l} \)

\( k_{inS} \) inactivation parameter for substrate-dependent enzyme inactivation in \( \frac{l}{mmol\cdot min} \)

\( K_m \) Michaelis-Menten constant in \( \frac{mmol}{l} \)

\( K_{mA} \) Michaelis-Menten constant of donor substrate A in \( \frac{mmol}{l} \)

\( K_{mB} \) Michaelis-Menten constant of acceptor substrate B in \( \frac{mmol}{l} \)

\( K_{mP} \) Michaelis-Menten constant of product P in \( \frac{mmol}{l} \)

\( \lambda \) event rate of a Poisson distribution

\( \mu \) expected rate of a Gaussian distribution

\( N \) Number of network nodes or communities

\( n \) degree of a network node or sample size

\( P \) fraction of sites belonging to the largest cluster (community)

\( p \) occupancy of the lattice sites or number of sequence pairs

\( p_c \) percolation threshold

\( p(d) \) number of sequence pairs with distance \( d \)

\( s \) community (cluster) size

\( s_\xi \) characteristic cluster size (community size)

\( \sigma \) either used as critical exponent for percolation (dependent on lattice geometry) or as standard deviation of a Gaussian distribution

\( t \) time

\( \tau \) Fisher’s exponent

\( \tau_h \) Fisher’s exponent derived from a histogram

\( \tau_{100} \) extrapolated Fisher’s exponent

\( \tau_{1/2} \) half-life time

22
Abstract

Enzymes are versatile catalysts for chemical reactions, mostly proteins, with various applications in white biotechnology due to their characteristic properties, such as reaction specificity, substrate specificity, regiospecificity or stereospecificity. Proteins are macromolecules composed of one or multiple chains of amino acids, with the amino acids serving as functional building blocks. In addition, cofactors, usually organic compounds or ions, can be required for an enzyme to catalyze chemical reactions. Thus, the function of an enzyme is determined by its molecular structure and composition, which is in turn encoded in the sequence of amino acids, the protein sequence.

The vast amount of data on protein sequences that is nowadays available in publicly accessible databases is a promising starting point for the exploration of new enzymatic functions or enzymes with improved catalytic characteristics. On the other hand, sophisticated techniques are required to organize these large data on protein sequences, to identify individual enzyme candidates in such large data and to relate protein sequences of such a candidate enzyme with homologous protein sequences of relevant function. Additionally, it is desirable to identify structurally equivalent amino acid positions across a family of homologous protein sequences, e. g. in order to prepare mutations for the design of enzymes with desired properties.

The first part of this thesis describes properties of protein sequence networks, which are generated by pairwise protein sequence alignments and are thereby a model for sequence-sequence relationships. Individual protein sequences are represented as nodes connected by weighted edges in a protein sequence network, with edges being defined by a threshold of pairwise sequence identity or sequence similarity. The concept of relating protein sequences in connected networks, rather than in strictly separated homologous protein families, allows the identification of individual protein sequences with interesting properties: The densely connected hubs of a protein sequence network hint at evolvable proteins, as a starting point for mutations in enzyme design. Protein sequences that are found as connectors between two different network communities, i. e. between two densely connected regions in a protein sequence network (equivalent of two homologous protein families), point at enzyme candidates with properties similar to both groups of enzymes. Furthermore, the concept of protein sequence networks allows to derive principles of enzyme evolution, such as the approximation of a fractal network dimension representing evolvability and likelihoods of amino acid mutations.

At the end of the first part, an exemplary protein sequence network of $\omega$-transaminases from fold type IV is used to highlight sequences that match characteristic amino acid positions or sequence motifs. The identification of these structurally equivalent positions is enabled by a newly implemented standard numbering scheme for fold type IV $\omega$-transaminases, showing sequence-structure relationships.
Besides bioinformatic analyses of protein sequence and structure, the biochemical characterization of enzyme candidates results in additional, heterogeneous data from various sources, which have to be linked with metadata on the conditions of the reaction or process under investigation. These biocatalytic data can be ratios, such as enantiomeric excess, yield or conversion of a biocatalytic reaction, or data on substrate or product concentrations over time, which are useful for investigating reaction kinetics. Hence, both bioinformatics and biocatalysis use various types of inhomogeneous data, including protein sequences, taxonomic information on source organisms, structural information on proteins, metadata on the reaction under investigation and experimental observations. It is desirable to unambiguously link these heterogeneous data with each other, which requires a strategy for data integration to facilitate both the search for new enzyme candidates and their experimental characterization.

The second part of this thesis describes implementations and applications of the BioCatNet database system, as a consistent approach to integrate the different data types mentioned above. The data model of the BioCatNet database system has been revised with regards to experimental data. The BioCatNet concept distinguishes between original data, such as protein sequences, structures and experimental data from collaborative research projects, and model-derived information, such as annotations of structurally equivalent positions within protein sequences or kinetic parameters that are derived from fitting kinetic models against measured time-course data.

The symmetric carboligation catalyzed by benzaldehyde lyase from *Pseudomonas fluorescens* is investigated as a test case for the BioCatNet database system. Data sets of time-courses for the two substrates benzaldehyde or 3,5-dimethoxybenzaldehyde are analyzed with different kinetic models, highlighting the influence of different models and their respective assumptions on the resulting kinetic parameter estimates. A kinetic model with substrate-dependent enzyme inactivation is found as the best fit for the data on both substrates.

The lessons learned from the effects of different kinetic model equations on the outcome of the parameter estimation allow in turn to refurbish the data management strategy of BioCatNet for future research projects on the exploration of enzymatic sequence-structure-function relationships. The concepts of protein sequence networks, standard numbering schemes and the estimation of kinetic parameters demand comparable requirements: The interpretation of the results from data mining approaches depend on both data and models, with the term model referring to concepts and assumptions that are applied on the original data, such as alignments that are applied on protein sequences, or kinetic models that are applied on time-course measurements. Therefore, both data and models should be combined in future data integration, to ensure reproducibility of data mining results, ideally in one common data- and modelbase.
Zusammenfassung

Enzyme sind vielseitige Katalysatoren für chemische Reaktionen, meistens Proteine, mit verschiedenen Anwendungen in der weißen Biotechnologie aufgrund ihrer charakteristischen Eigenschaften, wie Reaktionsspezifität, Substratspezifität, Regioselektivität oder Stereospezifität.

Proteine sind Makromoleküle, welche aus einer oder mehreren Ketten von Aminosäuren aufgebaut sind, wobei die Aminosäuren als funktionelle Bausteine dienen. Zusätzlich können Kofaktoren, für gewöhnlich organische Verbindungen oder Ionen, erforderlich sein, damit ein Enzym chemische Reaktionen katalysieren kann. Folglich ist die Funktion eines Enzyms durch seine molekulare Struktur und Zusammensetzung bestimmt, welche wiederum in der Sequenz von Aminosäuren codiert ist, der Proteinsequenz.


Am Ende des ersten Teiles wird ein Beispielhaftes Proteinsequenznetzwerk für \( \omega \)-Transaminasen aus Faltungstyp IV verwendet, um Sequenzen hervorzuheben, die zu charakteristischen Aminosäurepositionen oder Sequenzmotiven passen. Das Auffinden dieser sich strukturell entsprechender Positionen wird durch ein neu implementiertes Standardnummerierungsschema für Faltungstyp-IV-\( \omega \)-Transaminasen ermöglicht, welches Sequenz-Struktur-Beziehungen aufzeigt.

Abgesehen von bioinformatischen Untersuchungen der Proteinsequenz und Struktur, führt die biochemische Charakterisierung von Enzymkandidaten zu weiteren, heterogenen Daten aus verschiedenen Quellen, welche mit Metadaten über die Bedingungen der untersuchten Reaktion oder des untersuchten Prozesses zusammengebracht werden müssen. Diese biokatalytischen Daten können Verhältnisse sein, wie Enantiomerenüberschuss, Ausbeute oder Umsatz einer biokatalytischen Reaktion, oder Daten über Substrat- oder Produktkonzentrationen über der Zeit, welche nützlich sind, um die Reaktionskinetik zu untersuchen. Dadurch verwenden sowohl die Bioinformatik als auch die Biokatalyse verschiedene Typen heterogener Daten, darunter Proteinsequenzen, taxonomische Information über Ursprungsorganismen, Strukturinformation über Proteine, Metadaten über die untersuchte Reaktion und experimentelle Beobachtungen. Es ist wünschenswert, diese heterogenen Daten eindeutig miteinander zu verknüpfen, was eine Strategie zur Datenintegration erfordert, um sowohl die Suche nach neuen Enzymkandidaten als auch ihre experimentelle Beschreibung zu erleichtern.


Die Lektionen, die aus den Einflüssen verschiedener kinetischer Modellgleichungen auf das Ergebnis der Parameterschätzung gelernt wurden, erlauben wiederum die Strategie zur Datenverwaltung von BioCatNet für zukünftige Forschungsprojekte über enzymatische Sequenz-
Struktur-Funktionsbeziehungen zu überarbeiten. Die Konzepte von Proteinsequenznetzwerken, Standardnummerierungsschemata und die Schätzung kinetischer Parameter verlangen vergleichbare Voraussetzungen: Die Interpretation der Ergebnisse aus den *Data-Mining*-Ansätzen hängt sowohl von den Daten als auch von den Modellen ab, wobei der Begriff Modell sich auf Konzepte und Annahmen bezieht, welche auf die Originaldaten angewandt werden, wie etwa Alignments, die auf Proteinsequenzen angewandt werden, oder kinetische Modelle, die auf Zeitverlaufsmessungen angewandt werden.

Daher sollten sowohl Daten als auch Modelle in einer zukünftigen Datenintegration kombiniert werden, um die Reproduzierbarkeit der *Data-Mining*-Ergebnisse zu gewährleisten, idealerweise in einer gemeinsamen Daten- und Modelbank.
1 Introduction

According to the IUPAC Compendium of Chemical Terminology\textsuperscript{1} enzymes are mostly proteins which catalyze chemical reactions, specific for a certain type of chemical reaction (reaction specificity), with only certain types of chemical compounds being converted (substrate specificity) at specific sites (regiospecificity). Furthermore, in case of chiral substrates, one of the enantiomers is converted preferably (stereospecificity).

Proteins are macromolecules, large polypeptides, composed of one or multiple chains of amino acids. Sometimes additional helper molecules named cofactors, which are usually organic compounds or ions, are necessary for an enzyme to catalyze a chemical reaction. Amino acids are functional building blocks encoding the molecular structure of an enzyme (or a protein in general) in the amino acid or protein sequence. The protein structure in turn determines the function of an enzyme. Related amino acid sequences are termed homologous sequences, conceptualized in protein families, and are expected to result in proteins of comparable structure and function. Accordingly, it is assumed that there exists a relationship between protein sequence, structure and function of enzymes. Mutations of amino acids are in turn expected to result in different structural or functional features of an enzyme. The dynamics of an enzyme-catalyzed reaction are described by kinetic model equations, representing the reaction velocity, i.e. the changes of substrate or product concentrations over time (Xie, 2013).

The biological and biochemical diversity observed in nature is expressed in the diversity of different data types, as descriptors of specific features and observations. Although this diversity may be found overwhelming, it is a promising and tempting thought to consistently integrate these different data and their aspects for subsequent routines of computational analyses (data mining).

This chapter describes available data sources for sequence, structure and function of enzymes, as well as concepts of sequence-sequence and sequence-structure relationships. Finally, enzyme function is described with a focus on enzyme kinetics, as outlined for the specific exemplary kinetic model equations at the end of this chapter.

1.1 Data on sequence, structure and function of enzymes

Protein sequences are usually stored in large, publicly accessible databases, such as the non-redundant protein database from the National Center for Biotechnology Information (NCBI) (Benson \textit{et al.}, 2018), the UniProt database (Bateman \textit{et al.}, 2017) or the DNA Data Bank of Japan (DDBJ) (Kaminuma \textit{et al.}, 2011). Due to advances in sequencing technologies, the number of known protein sequences in these public databases is steadily increasing, raising the

ne for data management to facilitate the navigation in these large sequence data sets, as it was outlined by the UniProt Consortium for its protein sequence databases (Bateman et al., 2017). One of the central databases in the UniProt environment, the UniProt Knowledge Base (KB), removed approximately 47,000,000 redundant entries in March 2015, still resulting in tens of millions of non-redundant sequence entries, which exemplifies challenges in managing these high amounts of data (Bateman et al., 2017). In addition to databases focused on protein sequences, the Pfam database (Finn et al., 2015) organizes data on protein families by profile hidden Markov models (HMMs), stochastic representations of sequence profiles derived from multiple sequence alignments (reviewed by Eddy, 1998).

Information on individual protein structures is available in the Protein Data Bank (PDB), with approximately 90% of all protein structures being resolved by crystallography (Burley et al., 2017). As an illustration, the UniProt database system currently contains tens of millions of sequence entries (Bateman et al., 2017), whereas the number of individual protein structures in the PDB is approximately $10^6$ (Burley et al., 2017). Data on protein sequences and structures have been proven as useful for the identification of enzymes with new functions (reviewed by Lobb and Doxey, 2016).

For the investigations of relationships between protein sequence, structure and function of enzymes, the different data sources mentioned above have to be considered in combination, which demands technical means of data integration. Such a starting point for later data analyses was provided by the previously developed data warehouse system for analyzing protein families (DWARF) that combined data on both protein sequence and protein structure in a common data model (Fischer et al., 2006). This concept was later extended to link data from biocatalytic experiments unambiguously to protein sequences of the respective enzymes, as outlined in the first concepts of the BioCatNet database system by Reusch (2014) and Vogel (2015), extending the data model of DWARF (Fischer et al., 2006) for time-course data from biocatalytic measurements. Besides that, the BioCatNet data model from Reusch (2014) and Vogel (2015) was also designed to contain more details on taxonomic data of source organisms or source names that are linked to individual sequence entries.

Furthermore, there exist comprehensive and publicly available repositories on enzyme function, such as the Braunschweig Enzyme Database (BRENDA) (Schomburg et al., 2012), ExplorEnz (McDonald et al., 2007), the Kyoto Encyclopedia of Genes and Genomes (KEGG) REACTION (Kanehisa et al., 2006), and SABIO-RK (Wittig et al., 2012), which use literature to collect biochemical data and information. In addition to these repositories, there have been detailed suggestions for the description of experimental data on enzymes, such as the ones provided by the Standards for Reporting Enzymology Data (STRENDAD) Consortium (Tipton et al., 2014)\(^2\), as well as additional recommendations for data on biocatalytic reactions.

---

1.2 Sequence-sequence relationships: networks

(Gardossi et al., 2010). The concepts of STRENGDA have recently been implemented for direct data submission from experimenters, as a validation of data consistency prior to publication (Swainston et al., 2018).

One type of investigations on enzyme function focuses on the kinetics of an enzyme-catalyzed reaction, i.e., the identification of model equations and corresponding parameter values that describe the changes in substrate or product concentration over time (Xie, 2013). These kinetic models can be helpful in the identification of optimized process conditions for an enzyme-catalyzed reaction (Vasić-Rački et al., 2003; Almquist et al., 2014). Often, enzyme kinetic studies focus on the initial rate of an reaction, missing information on the complete time-course (also termed progress curve data) of substrate depletion or product formation, whereas time-course data can provide deeper insights into the dynamics of a reaction system (Duggleby and Morrison, 1978; Duggleby, 2001).

It would be desirable to combine the different resources on protein sequences and structures mentioned above with experimental data. Furthermore, one could think of computational tools to

1. identify enzyme candidates in large sequence data sets
2. describe evolutionary constraints for enzyme sequences
3. identify structurally equivalent positions within an enzyme family
4. analyze multiple sets of annotated time-course data by various models

1.2 Sequence-sequence relationships: networks

Atkinson et al. (2009) were among the first to apply the concept of networks (also named graphs) on comprehensive data sets of protein sequences in an approach named Sequence Similarity Network (SSN), based on the heuristics of BLAST (Basic Local Alignment Search Tool) for pairwise combinations within a sequence set and the choice of an E-value threshold to define edges in a network. The E-value describes the expectation value of finding random matches during the local alignments of a BLAST search against a sequence database (Altschul et al., 1990).

In such a network representation, protein sequences are regarded as nodes connected by weighted edges, which in turn indicate a metric of sequence homology. Sequence similarity networks constructed by the approach from Atkinson et al. (2009) were proven helpful in the investigations of various sets of sequences, such as in recent studies on tautomerases (Davidson et al., 2018) or proteins from Geminiviridae (Vaghi Medina et al., 2017). In this work, networks between homologous sequences were constructed differently, as outlined in chapters A.1 and
A.2, and were thus named more generally protein sequence networks, instead of the term sequence similarity networks that was coined for the approach from Atkinson et al. (2009). Here, the alignment techniques from the heuristics of USEARCH (Edgar, 2010) or pairwise alignments according to Needleman and Wunsch (1970) were used to construct networks, instead of the BLAST-based SSN approach from Atkinson et al. (2009).

Atkinson et al. (2009) also stated that a network representation of pairwise distances between protein sequences is generally advantageous in comparison to phylogenetic trees, since the latter projects relationships from multiple sequence alignments onto a single dimension, whereas networks are able to reveal multiple neighborhood relationships between protein sequences, thereby possibly showing links between divergent subfamilies of sequences that would be omitted in a phylogenetic tree (Atkinson et al., 2009). Since phylogenetic trees are derived from multiple sequence alignments, in contrast to protein sequence networks which are derived from pairwise alignments of sequences against each other, they are not able to capture relationships of protein sequences that are dissimilar to most of the other sequences in an alignment (compare with Rost, 1999).

For the protein family of TEM β-lactamases, protein sequence networks could be constructed with edges defined as mutations of single amino acids, due to the high microdiversity for this protein family, thereby following an alternative approach for the construction of protein sequence networks (Zeil et al., 2016)

1.2.1 Exemplary properties of networks

TEM β-lactamases were shown to form a scale-free network (Zeil et al., 2016). A scale-free network results in a typical degree distribution, with the term degree referring to the number of adjacent nodes of a given node from the network\(^3\).

The typical degree distribution of a scale-free network is described by a power law

\[
N(n) \sim n^{-\gamma},
\]

with \(N\) as the number of degrees \(n\) and a scaling exponent \(\gamma\), describing a "small world" network with few highly connected nodes (reviewed by Barabási and Albert, 1999). Such power law distributions have been observed in various environments, e. g. metabolic rates (West and Brown, 2004), population genetics (Manhart et al., 2012) or sequence networks of proteins (Enright et al., 2003), interpreting the power law distribution as a consequence of evolution (Deeds et al., 2003; Koonin et al., 2002) or protein structure (Deeds et al., 2003; Wuchty, 2001). Furthermore, it was proposed that scale-free networks possess a rather robust network topology, due to the connectivity of few network hubs (Albert et al., 2000). Power

\(^3\) The number of neighboring nodes is also named the degree centrality of a node.
law distributions of scale-free networks have been mentioned frequently, but sampling of the data can result in misleading conclusions on the actual network properties (Fox Keller, 2005; Lima-Mendez and van Helden, 2009). Even if selected subsets of a network result in a degree distribution following a power law, their scaling exponent might differ from the larger network (Stumpf et al., 2005). Thus, alternative distributions besides a power law have to be tested on various data and subsets of these data, to underline the hints for a scale-free network.

As a further network property, the fractal network dimension $D_f$ has been described as a parameter of the network geometry (reviewed by Saberi, 2015). The concept of a fractal network dimension can be linked to percolation theory which models transitions between communities\(^4\), i.e. transitions between clusters of a lattice model representing a network (Saberi, 2015; Fisher, 1967). In chapters 2.1.2 and A.2, the concept of percolation theory is applied to protein sequence networks by investigating the number $N(s)$ of communities with given community size $s$ (the number of sequences belonging to a homologous group of sequences).

Percolation theory describes the cluster distribution on a randomly populated lattice, with a parameter $p$ describing the occupancy of the lattice sites (Christensen and Moloney, 2005). For increasing values of $p$, the characteristic cluster size $s_\xi$ and the fraction $P$ of sites belonging to the largest cluster increases. As $p$ approaches the percolation threshold $p_c$, an infinite cluster appears for the first time on an infinite lattice, while on a finite-sized lattice the largest cluster percolates between the lattice boundaries. The core of percolation theory is a set of scaling relations that depend on $|p_c - p|$, such as

\[
s_\xi \sim |p_c - p|^{-1/\sigma} \quad (1.2)
\]

\[
P \sim (p - p_c)^\beta \quad (1.3)
\]

with critical exponents $\sigma$ and $\beta$ that depend on the geometry of the lattice. Most importantly, percolation theory predicts that the cluster size distribution $N(s)$ (the number $N$ of clusters with size $s$) decreases for $s \ll s_\xi$ as $N(s) \sim s^{-\tau}$ and decays exponentially for $s \gg s_\xi$. Near to percolation ($p \rightarrow p_c$), $s_\xi$ becomes infinite. Thus, for $s$ spanning many orders of magnitude, log $N(s)$ depends linearly on log $s$, with Fisher’s exponent $\tau$ describing the ratio of small to large clusters in the log-log plot of a histogram (Fisher, 1967).

1.3 Sequence-structure relationships: standard numbering schemes

Standard numbering schemes allow the identification of equivalent sequence positions across members of a protein family by assigning unambiguous reference position numbers, as it was previously shown for protein families such as the class B $\beta$-lactamases (Garau et al., 2004)\(^4\) Communities of a network are also named subgraphs or clusters.
or thiamine diphosphate-dependent decarboxylases (Vogel et al., 2012). Standard numbering schemes require a consistent multiple sequence alignment that is supposed to represent the protein family under investigation. In order to identify structurally equivalent positions within a protein family, a structural superimposition of representative structures is used to derive a structure-based multiple sequence alignment, i.e. an alignment in which the respective columns represent the structural superimposition (compare with Vogel et al., 2012; Russell and Barton, 1992). The commercially available 3DM databases follow a similar approach by using multiple structure alignments of reference enzymes to describe an enzyme family (Kuipers et al., 2010b). The reference alignment can be refined manually to consistently align secondary structure elements, especially in loop-rich regions (compare with Vogel et al., 2012). One protein sequence from the reference alignment is selected to assign position numbers to all remaining protein sequences from the respective protein family. To facilitate the alignment process, a profile HMM is derived as stochastic representation of the multiple sequence alignment (reviewed by Eddy, 1998). All sequences from a protein family are aligned against the reference profile HMM and the position numbers from the selected reference sequence are assigned accordingly.

1.4 Exemplary enzymes and enzyme families

1.4.1 \( \omega \)-transaminases

Transaminases (TAs) catalyze the transfer of an amino group from an amine donor to a carbonyl acceptor, using pyridoxal 5'-phosphate (PLP) as cofactor (Jansonius, 1998). Besides transaminases, there are also further enzymes using PLP as cofactor such as lyases, oxidoreductases and hydrolases (Percudani and Peracchi, 2009). There are two subgroups of TAs: \( \alpha \)-transaminases transfer an amino group exclusively to a carbonyl group in \( \alpha \)-position to a carboxyl group, in contrast to \( \omega \)-transaminases (oTAs or \( \omega \)-TAs) that show a wider spectrum of possible acceptor or donor substrates (Braunstein, 1973).

The case of transaminases shows the difficulties in distinguishing subfamilies of enzymes, since there are different principles and assumptions for the annotation of groups within the transaminase enzyme family, namely protein sequence, structure and function:

1. Sequence classification assigns TAs to five different aminotransferase classes based on sequence similarity (Jensen and Gu, 1996; Ouzounis and Sander, 1993; Mehta et al., 1993; Lyskowski et al., 2014; Grishin et al., 1995). Alternatively, TAs were sorted into six Pfam groups (Rausch et al., 2013; Finn et al., 2015), with phylogenetic trees sorting \( \alpha \)-TAs and \( \omega \)-TAs in different Pfam groups (Rudat et al., 2012; Arvidsson et al., 2001).

2. Structure-based classification sorts \( \alpha \) and \( \omega \)-TAs into two different types of protein folds (fold type I and IV) (Grishin et al., 1995; Schneider et al., 2000; Eliot and Kirsch, 2004;
1.4 Exemplary enzymes and enzyme families

Pavkov-Keller et al., 2016).

3. Functional classification distinguishes between α-TAs and ω-TAs (Ouzounis and Sander, 1993).

In addition to existing classification schemes on the levels of sequence, structure and function, Höhne et al. (2010) applied characteristic positions and motifs of protein sequences to predict subgroups with respective enzyme activity for fold type IV such as: 4-amino-4-deoxychorismate lyase (ADCL), (R)-amine transaminase (ATA), L-branched chain amino acid aminotransferase (BCAT) and D-amino acid aminotransferase (DATA). It would be desirable to combine the information on protein sequence and structure of transaminases in order to identify structurally equivalent sequence positions.

1.4.2 Benzaldehyde lyase: a ThDP-dependent decarboxylase

Thiamine diphosphate (ThDP)-dependent enzymes catalyze various chemical reactions, such as ligation or cleavage of carbon-carbon, carbon-nitrogen, carbon-sulfur or carbon-oxygen bonds by using the cofactor ThDP as a catalyst (reviewed by Hailes et al., 2013; Kluger and Tittmann, 2008). From all these different reactions, cleavage or formation of carbon-carbon bonds has gained most attention (Müller et al., 2013). By differences in the arrangement of protein domains, ThDP-dependent enzymes can be grouped into different superfamilies, such as the ThDP-dependent decarboxylases (Duggleby, 2006; Vogel and Pleiss, 2014).

One of these ThDP-dependent decarboxylases is benzaldehyde lyase (BAL), which was first discovered in Pseudomonas fluorescens Biovar I, allowing the strain to grow on benzoin as carbon source (González and Vicuña, 1989). BAL from P. fluorescens (PfBAL) cleaves the chiral hydroxyketone (R)-benzoin forming benzaldehyde (BA), but it also catalyzes the symmetric carboligation (also termed self-ligation) from benzaldehyde to benzoin (Figure 1.1), among other asymmetric C-C bond formations (Janzen et al., 2006). The enantioselective synthesis of hydroxyketones catalyzed by PfBAL was investigated in more detail due to its commercial relevance, as chiral hydroxyketones are building blocks for drugs (Demir et al., 2001).

For kinetic modeling of the symmetric carboligation catalyzed by PfBAL, the alternative substrate 3,5-dimethoxy-benzaldehyde (DMBA) was monitored by a fluorimeter approach in the work of Zavrel et al. (2008) (Figure 1.1). Similarly, the reversible reaction from benzaldehyde to benzoin was recently investigated by various kinetic models in the work of Ohs et al. (2018). Both Zavrel et al. (2008) and Ohs et al. (2018) tested different kinetic model equations for the ordered bi-uni reaction mechanism of the symmetric carboligation catalyzed by PfBAL.
1.5 Kinetic model equations for symmetric carboligation

For the type of reaction catalyzed by benzaldehyde lyase from *P. fluorescens*, it was shown earlier by Dünkelmann et al. (2002) that one molecule of substrate is a donor and the other an acceptor substrate. Whereas previous work on the kinetics of *Pf* BAL-catalyzed reactions assumed a single Michaelis-Menten parameter $K_m$ (Stillger et al., 2006; Hildebrand et al., 2007), the work from Zavrel et al. (2008) applied models with two respective $K_m$ parameters to capture the binding of both donor and acceptor substrate separately (Figure 1.2).

The symmetric carboligation catalyzed by *Pf* BAL serves as a test case for comparing the effect of different kinetic model equations on the resulting parameter estimates (Manuscript 5). The reactions under investigation comprise the conversion of benzaldehyde (BA) (Ohs et al., 2018) and 3,5-dimethoxy-benzaldehyde (DMBA) (Zavrel et al., 2008) to the respective products benzoin and 3,3',5,5'-tetramethoxy-benzoin (Figure 1.1).

In the following, the term kinetic model refers to the mathematical representation of a reaction mechanism. Thus, one reaction mechanism can be described by more than one set of mathematical equations, depending on the assumptions for the kinetic model. Following the naming convention used in (Zavrel et al., 2008), macrokinetic models are distinguished from microkinetic models. The term macrokinetic model refers to kinetic model equations which summarize elementary rate parameters in macrokinetic parameters, such as the Michaelis-Menten parameter $K_m$ that can be expressed by elementary rate parameters $k_1$, $k_\text{\textunderscore}1$ and $k_3$:
1.5 Kinetic model equations for symmetric carboligation

**Model 1** is an irreversible macrokinetic Michaelis-Menten model, with adapted stoichiometry taking into account two identical substrates,

\[
\frac{dc_P}{dt} = -\frac{1}{2} \frac{dc_A}{dt} = \frac{k_{catf} c_{Et} c_A}{K_{mA} + c_A},
\]

with time \( t \), substrate concentration \( c_A \), product concentration \( c_P \) and total enzyme concentration \( c_{Et} \). **Equation 1.5** has two independent parameters, the rate constant of the forward reaction \( k_{catf} \) and the macrokinetic Michaelis-Menten parameter \( K_{mA} \), summarizing the elementary rate parameters according to **Equation 1.4**.

**Model 2** is a macrokinetic representation of the (reversible) ordered bi-uni reaction mechanism, where two identical molecules of substrate are ligated to form one product (**Figure 1.2**).
1.5 Kinetic model equations for symmetric carboligation

The ordered bi-uni mechanism comprises two events for substrate binding, the first binding of a donor substrate molecule and the second binding of an acceptor substrate molecule. Model 2 is formulated as

\[
\frac{dc_P}{dt} = -\frac{1}{2} \frac{dc_A}{dt}
\]

\[
\frac{dc_P}{dt} = \frac{k_{catf}}{K_{iA} K_{mB}} \left( c_A^2 - c_P \right) \frac{c_{Ef}}{K_{eq}}
\]

with relations

\[
K_{iA} = K_{mB} - K_{mA}
\]

\[
K_{eq} = \frac{c_P}{c_A}
\]

\[
K_{mP} = \frac{K_{mB} (K_{mB} - K_{mA})^2 K_{eq}}{2K_{mA} K_{iA}}
\]

\[
K_{iB} = \frac{K_{mB} K_{iA}}{K_{mA} \left( 1 - \frac{K_{mA}}{K_{iA}} - 1 \right) \frac{K_{mP}}{K_{eq} K_{mB} K_{iA}}}
\]

and four model parameters: the rate constant of the forward reaction \( k_{catf} \), the equilibrium constant \( K_{eq} \) (see Equation 1.9), and the Michaelis-Menten parameters of the respective donor and acceptor binding (\( K_{mA} \) and \( K_{mB} \), respectively). The dependent parameters are the Michaelis-Menten parameter of the product in the reverse reaction (Equation 1.10) and the inhibition constants of the donor and acceptor substrate (Equations 1.8 and 1.11, respectively) which are derived from the model parameters as in Zavrel et al. (2008). The relations in Equations 1.8 and 1.9 result from the simplifying assumption that the elementary rate parameters are identical for both donor and acceptor substrate, i.e. \( k_1 = k_2 \) and \( k_-1 = k_-2 \).

Model 3 is the microkinetic equivalent of model 2 in Equations 1.6 to 1.11 and thus a different mathematical representation for the ordered bi-uni mechanism:

\[
\frac{dc_P}{dt} = -\frac{1}{2} \frac{dc_A}{dt}
\]

\[
\frac{dc_P}{dt} = \frac{k_1^2 k_3 c_A^2 - k_-1 k_-3 c_P}{k_-1 + k_-3 k_3 + k_1^2 c_A^2 + k_1 k_-3 c_A c_P + (2 k_1 k_3 + k_1 k_-1) c_A + 2 k_-1 k_-3 c_P}
\]

where the four independent parameters \( k_1, k_-1, k_3 \) and \( k_-3 \) represent the corresponding elementary reaction steps of the ordered bi-uni mechanism (Figure 1.2), assuming \( k_1 = k_2 \) and \( k_-1 = k_-2 \), too. As a consequence, the microkinetic parameters from model 3 can be
1.6 Data integration and data mining

converted to their macrokinetic counterparts and vice versa. The microkinetic parameters can be derived from the macrokinetic parameters of model 2 (Equation 1.7) by

\[ k_1 = k_2 = \frac{k_{catf}}{K_{mA}} \]  

\[ k_{-1} = k_{-2} = \frac{k_{catf}(K_{mB} - K_{mA})}{K_{mA}} \]  

\[ k_3 = k_{catf} \]  

\[ k_{-3} = \frac{k_{catf}}{K_{eq}(K_{mB} - K_{mA})} \]

and the macrokinetic parameters can be derived from the microkinetic parameters by

\[ k_{catf} = k_3 \]  

\[ K_{mA} = \frac{k_3}{k_1} \]  

\[ K_{mB} = \frac{k_{-1} + k_3}{k_1} \]  

\[ K_{eq} = \frac{k_3^2}{k_{-1}^2 k_{-3}} \]

**Model 4** is an extension of model 3 by substrate-dependent enzyme inactivation, i.e. it adds an additional term

\[ \frac{dc_E}{dt} = -k_{inS} c_A c_E, \]

with the inactivation parameter \( k_{inS} \) leading to five independent parameters in total: \( k_1, k_{-1}, k_3, k_{-3} \) and \( k_{inS} \).

1.6 Data integration and data mining

As Stitt and Gibon (2014) outlined for the example of systems biology, data on a single biological level are usually not informative at all levels, which points at the necessity of data integration. Investigations on protein sequence, structure and function of enzymes make use of various data types, ranging from simple text files for sequence information to various file types for experimental data. To investigate sequence-structure-function relationships of enzymes, these different data have to be combined by techniques of data integration, especially in case of high amounts of data. Advanced analyses of data are often named *data mining*, referring to the combination of multiple steps for data processing and analysis, following the previous steps of data integration, which is usually supported by databases. A database allows to write, read
1.6 Data integration and data mining

and delete data according to a data model, which is defined by the technical requirements of the underlying database management system. Such a data model is in turn designed by the requirements of a real-world application. One of the most frequently used data models is the relational data model that organizes data in unambiguously connected tables (Codd, 1970).

Data integration requires common standards for annotation, as it was exemplified by Gomez-Cabrero et al. (2014) for genomics data. Standards for data sharing are also required to ensure reproducibility of both data mining and experiments in general (Lapatas et al., 2015). The suggestions of the STREND Consortium (Tipton et al., 2014) and the requirements for reporting data on biocatalytic reactions by Gardossi et al. (2010) mentioned above are examples of such standards for data sharing. In addition, there exist even more generally applicable rules for data management, such as the FAIR data principles (Wilkinson et al., 2016). The FAIR data principles aim at making data findable (e. g. by assignments of metadata), accessible, interoperable (e. g. by usage of standardized vocabularies and data formats) and reusable (e. g. by community standards such as the STREND guidelines) (Tipton et al., 2014; Wilkinson et al., 2016).
2 Results

This thesis comprises analyses on different data types relevant to understanding the relationships between protein sequences, structural information and function of enzymes, with the latter focusing on enzyme kinetics. In previous projects and theses, an in-house database system for protein families had been established in the Bioinformatics research group of Prof. Dr. Jürgen Pleiss (University of Stuttgart, Germany). This database system serves as starting point for multiple routines of data mining by integrating data on protein sequences and structural information.

Chapter 2.1 comprises bioinformatic and statistic assessments of large data sets on protein sequences, as a starting point for the elucidation of sequence-structure-function relationships. In chapter 2.1.1 protein sequence networks were investigated with respect to highly connected nodes and approximations of network dimensions (Manuscript 1). The subsequent chapter 2.1.2 describes investigations on the connectivity of homologous families in protein sequence networks (Manuscript 2). The finding that newly discovered protein sequences often appear connected to already known homologous families could be shown in an additional study for different esterases (Manuscript 6).

A further type of database application is the setup of a family-specific protein database to link protein sequences and structural information followed by sequence comparisons with standard numbering schemes, which was exemplified for the family of \( \omega \)-transaminases (Manuscript 3). Furthermore, a standard numbering scheme was also implemented for the superfamily of transketolases (Manuscript 7).

Whereas the original intention of the BioCatNet database system was to provide means of data archival similar to an electronic laboratory journal, this thesis changes the scope of managing experimental data towards the analysis of experimental data by different kinetic models. The BioCatNet database system was implemented and tested for biocatalytic measurements (Manuscript 4). In addition, the BioCatNet data model was refurbished to describe more details on the reaction conditions and to allow for other types of enzymatic data besides time-course measurements, such as yield, conversion and enantiomeric excess of an enzymatic process. Test cases for the estimation of kinetic parameters from time-course data of an enzymatic reaction are described in (Manuscript 5) suggesting a revised data management strategy for both measured data and data derived from kinetic model equations.

2.1 Properties of protein sequence networks

As it is rare to find a protein family whose sequences differ mostly by single mutations, individual point mutations are not an applicable metric for pairwise distances between sequences from most protein families. There are, however, exceptions such as the family of TEM \( \beta \)-lactamases
2.1 Properties of protein sequence networks

which share a high microdiversity that in turn allows to construct protein sequence networks by point mutations of amino acids (see also Zeil et al., 2016). Therefore, protein sequence networks were formed by pairwise alignments of amino acid sequences from a respective protein family, as outlined in chapters A.1.3 and A.2.3. In the following, protein sequence networks were investigated with respect to the distributions of degree and community size, with the former being used for an approximation of network dimensions, too.

2.1.1 Distributions of degrees and network dimensions

In previous work by Zeil et al. (2016), a network of point mutations from the TEM β-lactamase core region, corresponding to TEM-1 positions 24 to 280, exhibited a degree distribution following a power law: The $N$ number of degrees (neighboring nodes) $n$ followed a power law distribution described by Equation 1.1 with a scaling exponent $\gamma \approx 1.2$ (Zeil et al., 2016). The family of TEM β-lactamases (TEMs) served as a test case for the comparison of a network of point mutations and a network based on pairwise sequence identity (Manuscript 1). A sufficiently high threshold of sequence identity was expected to result in a network topology equivalent to the point-mutation network. By applying a threshold of 99.5% pairwise sequence identity, a network of 267 nodes and 401 edges was formed in analogy to the point mutation network showing a degree distribution with a scaling exponent $\gamma$ of 1.2, too (Figure 2.1). The scaling exponent $\gamma$ was determined by linear regression in log-log space. Alternative distributions such as a Gaussian and a Poisson distribution were not able to describe the observed degree distribution (Figure A.7).

![Figure 2.1: Distribution of the number of sequences $N$ in a distance-based network of TEM β-lactamases having $n$ first neighbors. The degree distribution follows a power law with exponent $\gamma = 1.2$. Compare with Figures A.6 and A.7 (see also Figures A.1 and A.2).](image-url)
It was a tempting hypothesis, whether protein sequence networks for other enzyme families of low microdiversity, with homologous sequences differing by several mutations, would result in qualitatively similar degree distributions of Equation 1.1 as it was verified for the TEM \( \beta \)-lactamase family. Furthermore, it would be recommendable to investigate protein families of different sample sizes and folds. As test cases, protein sequence networks were formed for the enzyme families of \( \beta \)-hydroxyacid dehydrogenases/ imine reductases (bHADs), thiamine diphosphate-dependent decarboxylases (DCs), \( \omega \)-transaminases (oTAs) and short-chain dehydrogenases/reductases (SDRs), which showed qualitatively different distributions of pairwise sequence identities (Figure A.3). Due to low micodiversity, the threshold of pairwise sequence identity was set to 95\% for the four enzyme families bHADs, DCs, oTAs and SDRs, in contrast to the higher threshold of 99.5\% that had been used for TEMs above. The respective protein sequence networks resulted in qualitatively similar degree distributions (Figure 2.2), with the distribution \( N(n) \) differing remarkably from a power law distribution for higher degrees \( n \), reflecting the poor statistical sampling of the known sequence space. Thus, the scaling exponents \( \gamma \) were only approximated for less connected nodes (with \( n < 70 \) (for oTA, SDR) or \( n < 50 \) (for bHAD, DC)) by linear regressions in log-log space (Table 2.1), resulting in scaling exponents \( \gamma \) ranging from 1.2 to 1.3 (Table 2.1). For protein sequence networks of bHADs, DCs, oTAs and SDRs, the power law distribution was also observed for decreasing global sequence identity thresholds to \( \geq 90 \% \), \( \geq 85 \% \), or \( \geq 80 \% \) (compare with Figures A.8 to A.10), with the scaling exponents \( \gamma \) decreasing to 0.9 to 1.1 with decreasing sequence identity thresholds. The power law distributions of degrees point at highly connected hubs, i.e. sequences with many homologous sequences (Table A.3).

Table 2.1: Overview of the analyzed protein family networks by number of nodes, edges and maximal degree (number of neighbors) for a 95\% sequence identity threshold, with average sequence length. The small family of TEM \( \beta \)-lactamases is shown as reference, due to its high microdiversity, with a threshold of 99.5\% sequence identity (a). The enzyme families comprise TEM \( \beta \)-lactamas (TEMs), \( \beta \)-hydroxyacid dehydrogenases/ imine reductases (bHADs), thiamine diphosphate-dependent decarboxylases (DCs), \( \omega \)-transaminases (oTAs) and short-chain dehydrogenases/ reductases (SDRs). Values for \( \gamma \) refer to the scaling exponents of the degree distributions from Figures 2.1 and 2.2. Values for \( D_f \) refer to the slope in Figure 2.3. Compare with Table A.2.

<table>
<thead>
<tr>
<th>Enzyme family</th>
<th>Nodes</th>
<th>Edges</th>
<th>Maximal degree</th>
<th>Average sequence length</th>
<th>( \gamma )</th>
<th>( D_f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEM(^a)</td>
<td>267</td>
<td>401</td>
<td>86</td>
<td>250</td>
<td>1.2</td>
<td>1.8</td>
</tr>
<tr>
<td>bHAD</td>
<td>17,020</td>
<td>148,188</td>
<td>259</td>
<td>320</td>
<td>1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>DC</td>
<td>24,880</td>
<td>309,635</td>
<td>266</td>
<td>580</td>
<td>1.1</td>
<td>0.7</td>
</tr>
<tr>
<td>oTA</td>
<td>79,987</td>
<td>1,444,727</td>
<td>381</td>
<td>460</td>
<td>1.2</td>
<td>0.9</td>
</tr>
<tr>
<td>SDR</td>
<td>81,680</td>
<td>838,743</td>
<td>312</td>
<td>300</td>
<td>1.3</td>
<td>1.0</td>
</tr>
</tbody>
</table>
2.1 Properties of protein sequence networks

Figure 2.2: Degree distributions for the protein families with low microdiversity from Table 2.1 with neighbors defined by ≥ 95% global sequence identity. The corresponding scale-free exponents $\gamma$ were derived from linear regression for degrees $\leq 50$ (bHAD, DC) or $\leq 70$ (oTA, SDR) and are summarized in Table 2.1. See also Figure A.4.

As a further property of protein sequence networks, the (fractal) network dimension $D_f$ was approximated by counting the number of sequence pairs $p(d)$ that differed by less than $d$% (100% - sequence identity in %) for $d = 2, 4, 6, \ldots$ (Figure 2.3). Thus, the dimension $D_f$ was derived from linear regressions in log-log space from a cumulative distribution of sequence pairs (representing degrees of network nodes) over pairwise sequence distances from the respective networks. For low values of pairwise sequence distance $d$ ($d \leq 10$, i.e. > 90% identity), $\log p(d)$ increased linearly with $\log d$, resulting in a network dimension $D_f$ between 0.7 and 1.0 for the four superfamilies with low microdiversity from Table 2.1. For increasing pairwise distances $d$, the network dimensions $D_f$ increased to $D_f$ between 3.5 and 4.5 for $30 \leq d \leq 70$ (Figure 2.4).

For the family of TEM β-lactamases, $D_f$ was estimated to 1.8 from the values at $d = 2$ and $d = 4$. Due to the high sequence identities of the members from the TEM β-lactamase family, only few sequence pairs were found with distances > 4% identity. The estimation of $D_f$ for the
2.1 Properties of protein sequence networks

point-mutation network of TEM β-lactamases by comparing the number of single and double mutants resulted in a higher value of $D_f = 4.0$ (Figure A.12). For higher distances beyond double mutants, the limited network size for TEMs resulted in an apparent decrease of $D_f$, and the analysis of double, triple, and quadruple mutants resulted in $D_f = 1.8$, as observed for the TEM β-lactamase network based on pairwise sequence identities (Figure A.5).

Figure 2.3: Cumulative distributions of sequence pairs $p(d)$ for pairwise distances of $d$ % for the protein families TEM (open squares), DC (filled squares), bHAD (open circles), SDR (diamonds) and oTA (filled circles) from Table 2.1 in subsequent distance intervals of 2 % distance $d$ (100 % - sequence identity in %). Linear fits are shown as red lines for distances up to 10 % identity (up to 4 % for TEM). For further distances between 30 and 70 %, an approximately linear area is depicted in red (compare with Figure 2.4). See also Figure A.5.
Figure 2.4: Detail view from Figure 2.3 for the protein families DC (filled squares), bHAD (open circles), SDR (diamonds) and oTA (filled circles) showing cumulative distributions of sequence pairs $p(d)$ for pairwise distances of $d \%$. Regression lines are given for the sequence pairs from the DC (blue line), bHAD (red line), SDR (blue dashed line) and oTA family (red dashed line).
2.1 Properties of protein sequence networks

2.1.2 Distributions of community sizes

The topology of the known sequence space was further analyzed for the four large protein superfamilies of low microdiversity from Table 2.1 by analysis of the community (or cluster) size distribution (Manuscript 2). By applying increasing thresholds of pairwise sequence identity, the number of edges in a protein sequence network decreases, leading to the occurrence of several isolated subgraphs, i.e. internally connected communities (clusters) which emerge without connections to other communities. For each of the four protein superfamilies SDR, oTA, DC and bHAD, sequences were clustered by an exemplary threshold of 60% global sequence identity. The number $N$ of communities with size $s$ was analyzed in a histogram with logarithmic bins for $s$ between 1 and 10, 11 and 100, 101 and 1,000, 1,001 and 10,000, and 10,001 and 100,000 to improve statistical sampling (Figure 2.5).

![Figure 2.5: Distribution of community sizes, resulting from subgroups clustered by 60% pairwise sequence identity determined by USEARCH (Edgar, 2010), for the protein families SDR, oTA, DC and bHAD following a power law distribution according to Equation 2.1. Compare with additional results in Figure A.14.](image-url)
The distribution of community sizes followed a power law

\[ N(s) \sim s^{-\tau}, \]  

which was observable by a linear dependency of \( \log s \) and \( \log N(s) \) for the four different protein superfamilies (SDR, oTA, DC and bHAD). The Fisher exponent \( \tau_h \) of a histogram describes the ratio between small and large communities and is derived from linear regression in the log-log plot of the histogram (Figure 2.5). From the Fisher exponent \( \tau_h \) of the histogram, the Fisher exponent \( \tau \) of the underlying community size distribution was calculated by fitting the observed \( \tau_h \) of the histogram to a model distribution of community sizes following a power law distribution\(^5\). Though the protein families differed in size, structure, and function, the values for the extrapolated Fisher exponent \( \tau_{100} \) varied only slightly (Table 2.2).

While the Fisher exponent \( \tau \) was almost independent of the protein family and its size, its absolute value depended on the threshold criterion used for clustering. Upon clustering of the four families with four thresholds between 60 and 90\%, the community size distributions followed a power law for all thresholds. Consequently, the Fisher exponent \( \tau \) increased almost linearly with increasing thresholds from \( \tau_{60} \) between 1.6 and 2.0 at 60\% threshold, to \( \tau_{90} \) between 2.2 and 2.9 at 90\% threshold (Figures 2.6 and A.17).

The Fisher exponent \( \tau \) was extrapolated to a threshold of 100\%, simulating a network of nodes separated by single mutations (\( \tau_{100} \)). For the four protein families, the extrapolated \( \tau_{100} \) values varied between 2.3 and 2.8 (Table 2.2). The Fisher exponent \( \tau \) was also observed to increase with a decreasing number of nodes, as it was observed for protein sequence networks of randomly selected \( \alpha/\beta \)-hydrolases (abHs) (Figure A.18). Additionally, many of the protein sequences used in Manuscript 6 were found to be homologous to sequence entries already present in the lipase engineering database (LED)\(^6\), suggesting that newly discovered proteins emerge rather within connected communities and less frequently at the periphery of a protein sequence network.

\(^5\) https://doi.org/10.1371/journal.pone.0189646.s006, accessed on April 30, 2018.

\(^6\) http://www.led.uni-stuttgart.de/, accessed on April 30, 2018.

Table 2.2: Protein superfamily size and the Fisher exponent extrapolated to 100\% sequence identity (\( \tau_{100} \)) of the four protein families. Compare with additional results in Table A.6.

<table>
<thead>
<tr>
<th>Enzyme superfamily</th>
<th>Superfamily size</th>
<th>( \tau_{100} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDR</td>
<td>14,100</td>
<td>2.4</td>
</tr>
<tr>
<td>oTA</td>
<td>121,000</td>
<td>2.3</td>
</tr>
<tr>
<td>DC</td>
<td>39,000</td>
<td>2.8</td>
</tr>
<tr>
<td>bHAD</td>
<td>31,000</td>
<td>2.5</td>
</tr>
</tbody>
</table>
2.1 Properties of protein sequence networks

Figure 2.6: Fisher exponent $\tau$ of the community size distributions of protein superfamilies from Figure 2.5 for clustering thresholds between 60 and 90\% pairwise sequence identity with extrapolated Fisher exponent $\tau_{100}$ determined by linear regression. Compare with additional results in Figure A.15.
2.1 Properties of protein sequence networks

2.1.3 Exemplary sequence networks of $\omega$-transaminases

Within this work, a family-specific protein sequence database was implemented for $\omega$-transaminases, the $\omega$-transaminase engineering database (oTAED) (Manuscript 3), by methods outlined in more detail in chapter A.3. Protein sequences in the oTAED were sorted into two superfamilies named fold types I and IV. If the majority of sequence entries within a homologous family in the oTAED were longer than 350 amino acids, the respective homologous family was assigned to the fold type I superfamily. Consequently, homologous families were assigned to the fold type IV superfamily, if most of their sequences were shorter than 350 amino acids. While protein sequences from fold type I showed an average length of 432 amino acids in 124 homologous families, protein sequences from fold type IV had an average length of 297 amino acids in 45 homologous families (Figure A.25). Most of the sequence entries for both fold type I and IV were sorted in one large homologous family, underlining the concept of highly connected protein sequence networks that was shown in the distributions of community sizes in chapter 2.1.2.

Standard numbering schemes were established for the two superfamilies fold type I and fold type IV as described in chapter A.3.3 (in an equivalent approach as for the transketolase standard numbering scheme from Manuscript 7). The respective standard numbering schemes for fold type I and IV were used to identify conserved positions and positions for substrate-specificity mentioned in literature within the oTAED, based on alignments of representative sequences (Table A.12). Furthermore, the Fold type IV standard numbering scheme was applied to search for protein sequences that matched sequence motifs or characteristic positions from Hölme et al. (2010) (Table A.13).

In addition, these sequence positions and motifs were used to identify matching sequences in a protein sequence network for representative fold type IV sequences (Figure 2.7), which was constructed as described in chapter A.3.3. The annotation of the protein sequence network showed that sequences matching positions and sequence motifs for ATA, ADCL, BCAT and DATA were in relation with ($R$)-selective $\omega$-TAs from fold type IV (Figure A.24). Besides that, several of the annotated sequences appeared to form separate communities within network.

Furthermore, protein sequence networks of fold types I and IV were also applied to identify nodes for sequences from extremophilic source organisms (Figure A.26). The information on the growth conditions of the respective source organism was retrieved from the BacDive database\(^7\), in combination with entries on taxonomic names and their possible synonyms from the oTAED. In contrast to the previous annotation of sequence motifs and characteristic positions for the fold type IV sequence network, the occurrence of extremophilic source organisms did not form communities within the sequence networks of fold types I and IV (Figure A.26).

\(^7\) https://bacdive.dsmz.de/, accessed on April 30, 2018.
2.1 Properties of protein sequence networks

Figure 2.7: Example of a protein sequence network for 447 representative sequences from the superfamily of $\omega$-transaminases from fold type IV, clustered by 30% sequence identity in USEARCH (Edgar, 2010) with 13,752 edges defined by a threshold of 50% sequence similarity (derived from pairwise alignments according to Needleman and Wunsch (1970). Sequences are annotated as 4-amino-4-deoxychorismate lyase (blue), (R)-amine transaminase (green), L-branched chain amino acid aminotransferase (red) or D-amino acid aminotransferases (black). See also Figure A.24.
2.2 Kinetic modeling in the BioCatNet database system

The BioCatNet database system was established as repository for the sequence-structure-function relationships of enzyme families (Manuscript 4), as it is described in chapter A.4. In case of protein sequences and structures, original data is parsed from publicly available data repositories, in contrast to measured data from biocatalytic experiments that are uploaded directly from experimenters within collaborative research projects, as in the case studies for kinetic modeling in chapter A.5 (Manuscript 5). The minimal requirements for data submission of experimental measurements to BioCatNet constitute a compromise between completeness and usability. Therefore, the BioCatNet data model distinguishes between mandatory and optional attributes, with complete annotations of reaction conditions being highly recommended to enhance reproducibility of both experimental investigations and computational analyses, since the latter cannot be interpreted without considering all possibly relevant factors (Table A.14). To facilitate the submission of biocatalytic data to BioCatNet, a formatted template of a Microsoft Excel spreadsheet is provided on the BioCatNet documentation website.

2.2.1 Extensions of the relational data model

The BioCatNet database system makes use of the database management system Firebird, allowing to manipulate the relational data model, i.e. the tables and relationships between tables for storage and linkage of data. Since not all research projects on enzymes are related to the field of kinetic modeling, additional types of experimental measurements should be included within the BioCatNet database system, to broaden to scope of possible applications beyond time-course data on substrate depletion or product formation. Therefore, the relational data model of BioCatNet was extended to allow for experimentally determined ratios, such as enantiomeric excess (ee), yield or conversion of an enzymatic reaction, as these data types often occur in screening experiments of enzyme candidates, which might be investigated with respect to kinetics in a further round of experiments (Figure A.27).

Whereas measurement data on time-courses are stored in a separate table of the BioCatNet data model (Reusch, 2014), an additional table was provided for measured ratios (ee, yield or conversion), named PARAMETER_MEASUREMENTS (Figure 2.8 and A.32). Since these ratios (measured parameters) could refer to a specific reference enzyme, such as a wild-type enzyme, a link to the SEQUENCES table was added, to allow for a linkage with the complete protein sequence of such a reference enzyme. Furthermore, the table for measured parameters (ee, yield or conversion) is linked with tables listing the chemical compound (i.e. substrate or product

---

2.2 Kinetic modeling in the BioCatNet database system

under investigation) and the measurement method (to allow for additional specifications on the measurement principle or the applied type of measurement device). An additional table named PARAMETER_LIST was added to make the data model adaptable to further parameters besides ee, yield or conversion.

Noteworthy, this part of the data model is not intended to store information on kinetic parameters, since their values are derived from fits of kinetic models equations against time-course data of concentrations over time (compare with Figure A.28). Thus, for investigations on enzyme kinetics, BioCatNet stores the original data from time-course experiments (chapter A.4), as it was shown for the data from Zavrel et al. (2008) and Ohs et al. (2018) on PfBAL-catalyzed symmetric carboligation.
2.2.2 Parameter estimations for PfBAL-catalyzed symmetric carboligation

The symmetric carboligation (also named self-ligation) catalyzed by benzaldehyde lyase from *Pseudomonas fluorescens* (PfBAL) was used as an exemplary reaction system to point out a data management strategy for biocatalytic experiments, with a focus on kinetic modeling.

Data sets for two substrates converted by PfBAL under comparatively similar conditions served as test cases: The first dataset comprised nine progress curves of 3,5-dimethoxybenzaldehyde (DMBA) depletion which were analyzed previously in the work of Zavrel *et al.* (2008). The second dataset comprised thirteen progress curves for the self-ligation of benzaldehyde (BA) to benzoin which were provided by the work from Ohs *et al.* (2018). The investigations of Zavrel *et al.* (2008) had identified a simplified microkinetic model for the self-ligation of DMBA, representing the bi-uni reaction mechanism (Equations 1.12 to 1.13). The term microkinetic model refers to the representation of elementary reaction steps, in contrast to the formulation of macrokinetic models which lump elementary reaction steps together, as used in Zavrel *et al.* (2008).

The microkinetic and macrokinetic parameter estimates from Zavrel *et al.* (2008) can be compared with each other, as both micro- and macrokinetic model have the same number of parameters, respectively. The proposed model simplification from Zavrel *et al.* (2008) assumes equal rate constants for the reversible binding of the first and the second substrate molecule of DMBA. In addition, the previous work from Ohs *et al.* (2018) identified an extended microkinetic model with substrate-dependent enzyme inactivation for the symmetric carboligation of BA.

It was tempting to see whether this model would also fit for the data of the DMBA-self-ligation from Zavrel *et al.* (2008). Furthermore, the results from Zavrel *et al.* (2008) and Ohs *et al.* (2018) were a motivation to compare the impact of different kinetic model equations on the resulting parameter estimates, such as the macro- and microkinetic model equations and the uncertainties of their respective parameter estimates. Besides that, these two datasets for the symmetric carboligation of BA and DMBA were chosen as a proof of principle for the concept of the BioCatNet data analysis workflow (Manuscript 5), since they provide heterogeneous data of experiments, with different sample sizes, reaction conditions and varying initial concentrations (Table A.15). While the data for DMBA were measured by a fluorimeter, thus providing a rather high sample size of 2786 data points (Zavrel *et al.*, 2008), the data for the symmetric carboligation of BA were measured by high performance liquid chromatography (HPLC), with a smaller sample size of 374 data points (Ohs *et al.*, 2018).

In the following, four different sets of kinetic model equations were applied for fits against both the DMBA and the BA data set, as described in the introductory chapter 1.5. The computational methods from Ohs *et al.* (2018) was applied as described in more detail in
2.2 Kinetic modeling in the BioCatNet database system

chapters A.5 and A.5.6.

**Application of the Michaelis-Menten model** Since the reactions investigated here are reversible (Figure 1.1 and 1.2), the irreversible two-substrate Michaelis-Menten model (model 1, Equation 1.5) cannot describe the complete time-course, leading to estimates for $K_{mA}$ and $k_{catf}$ with high uncertainty (Table A.18).

Therefore, the Michaelis-Menten model was applied to the first six minutes of the time-courses only, similar to initial rate measurements under presumably low product concentrations. For both the DMBA and BA dataset, the shortened time-courses could be fitted more reliably by the Michaelis-Menten model (as seen in the lower values for $\frac{RSS}{n}$), resulting in parameter estimates for both $K_{mA}$ and $k_{catf}$ with remarkably reduced relative standard deviations (Tables 2.3 and 2.4), in contrast to the fits for the complete time-courses.

The evaluation of the first six minutes of the reaction progress covers DMBA concentrations between 1.5 and 3 mM and BA concentrations between 30 and 75 mM (Table A.15), which is below and above the respective estimates for $K_{mA}$ of 2.6 mM and 55 mM from the fits with the Michaelis-Menten model. While the estimates for $K_{mA}$ differed by an order of magnitude, the estimated values for $k_{catf}$ were found similar for both DMBA and BA.

**Effects of model 1 compared to model 2** The resulting estimated kinetic parameters of the Michaelis-Menten model (model 1) covering the shortened time-courses (the first six minutes of the actual measurements) were compared to the estimates from the previously published macrokinetic model for the bi-uni mechanism (model 2) from Zavrel et al. (2008) applied to the complete time-courses (Figures A.38 to A.41). While the estimated values for $k_{catf}$ for both the BA and DMBA data set were in a comparable order of magnitude to the results for the Michaelis-Menten model (the latter applied to the shortened progress curve data only), the estimates for $K_{mA}$ were similar with model 1 for BA, but differed by up to four orders of magnitude for DMBA. Furthermore, the values of both $K_{mA}$ and $K_{mB}$ from model 2 showed larger relative standard deviations ($> 100\%$) implying that both $K_m$ parameters were unidentifiable for model 2, in contrast to the $K_m$ returned by the fit of model 1 for the initial time-courses.

**Effects of model 2 compared to model 3** As expected, the macrokinetic (model 2) and the microkinetic model (model 3) for the (reversible) ordered bi-uni reaction mechanism resulted in comparable values for $\frac{RSS}{n}$ for both DMBA and BA, which is also seen in the similar overall fit quality (Figures A.40 to A.43). In case of DMBA, the estimates for $k_{catf}$ (which is equivalent to $k_3$, Equation 1.16), and $K_{eq}$ were found identical, whereas estimates for both $K_{mA}$ and $K_{mB}$ differed approximately by an order of magnitude ($1.7 \cdot 10^{-4}$ and $2.9 \cdot 10^{-3}$ mM for $K_{mA}$,
2.2 Kinetic modeling in the BioCatNet database system

4.2 · 10^{-3} and 2.0 · 10^{-2} mM for \( K_{mB} \), respectively). Besides that, the microkinetic parameters \( k_1 \) and \( k_{-3} \) differed for models 2 and 3, too. Noteworthy, most relative standard deviations for the parameter estimates were remarkably smaller for the microkinetic model 3 than for the macrokinetic model 2, although both models describe the same ordered bi-uni reaction mechanism. In case of the BA data set, all estimated kinetic parameters differed by less than a factor of two. The low \( K_m \) values for DMBA and the high \( K_{eq} \) for BA gave doubts that models 2 and 3 were sufficient to describe the observed data.

**Effects of model 3 compared to model 4** For both substrates, the \( \frac{RSS}{n} \) and the relative standard deviations of the kinetic parameters were lower for the microkinetic model with substrate-dependent inactivation (model 4, as an extension of model 3), indicating a higher fit quality and more reliably identified parameter estimates in comparison to the remaining kinetic models (models 1 to 3) (Figures 2.9 and 2.10). Unexpectedly, the equilibrium constant differed by a factor of twenty between both DMBA and BA. The catalytic activity of PfBAL towards both DMBA and BA differed by less than a factor of two, whereas \( K_{mA} \) towards DMBA was more than two orders of magnitude lower than \( K_{mA} \) towards BA (Tables 2.3 and 2.4). The differences in \( K_{mA} \) and \( K_{mB} \) between the substrates DMBA and BA correlate to the two orders of magnitude difference in \( k_1 \) and \( k_{-3} \) between the substrates, whereas \( k_{-1} \) and \( k_3 \) were estimated similarly for both substrates. In addition, the inactivation parameter \( k_{mS} \) for BA was one order of magnitude smaller than the respective estimate of DMBA.
2.2 Kinetic modeling in the BioCatNet database system

Table 2.3: Estimated kinetic parameters and the residual sum of squares (RSS) over sample size $n$ for the benzaldehyde lyase-catalyzed self-ligation of 3,5-dimethoxy-benzaldehyde and four kinetic models\(^1,2\). Estimated parameter values are given in bold, whereas parameter values obtained from Equations 1.14 to 1.21 are set in regular font. The relative standard deviations (in %) are given in parentheses. Relative standard deviation values $< 100\%$ are marked with $*$ to indicate the higher reliability of the respective kinetic parameter values.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Model 1 $^3$</th>
<th>Model 2 $^4$</th>
<th>Model 3 $^5$</th>
<th>Model 4 $^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$ mM$^{-1}$ min$^{-1}$</td>
<td>-</td>
<td>$1.6 \cdot 10^7$</td>
<td>$9.3 \cdot 10^5$</td>
<td>$4.5 \cdot 10^4$</td>
</tr>
<tr>
<td>$k_{-1}$ min$^{-1}$</td>
<td>(501)</td>
<td>(108)</td>
<td>(14)$^*$</td>
<td></td>
</tr>
<tr>
<td>$k_3$ min$^{-1}$</td>
<td>(590)</td>
<td>(53)$^*$</td>
<td>(6)$^*$</td>
<td></td>
</tr>
<tr>
<td>$k_{-3}$ mM$^{-1}$ min$^{-1}$</td>
<td>-</td>
<td>$2.7 \cdot 10^3$</td>
<td>$2.7 \cdot 10^3$</td>
<td>$2.8 \cdot 10^3$</td>
</tr>
<tr>
<td>$k_{inS}$ mM$^{-1}$ min$^{-1}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>$8.0 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>$k_{catf}$ min$^{-1}$</td>
<td>$4.4 \cdot 10^3$</td>
<td>$2.7 \cdot 10^3$</td>
<td>$2.7 \cdot 10^3$</td>
<td>$2.8 \cdot 10^3$</td>
</tr>
<tr>
<td>$K_{mA}$ mM</td>
<td>$2.6$</td>
<td>$1.7 \cdot 10^{-4}$</td>
<td>$2.9 \cdot 10^{-3}$</td>
<td>$6.2 \cdot 10^{-2}$</td>
</tr>
<tr>
<td>$K_{mB}$ mM</td>
<td>-</td>
<td>$4.2 \cdot 10^{-3}$</td>
<td>$2.0 \cdot 10^{-2}$</td>
<td>0.2</td>
</tr>
<tr>
<td>$K_{eq}$ mM$^{-1}$</td>
<td>-</td>
<td>$3.9$</td>
<td>3.9</td>
<td>4.1</td>
</tr>
<tr>
<td>RSS $^n$ mM$^2$</td>
<td>$6.7 \cdot 10^{-3}$</td>
<td>$1.4 \cdot 10^{-3}$</td>
<td>$1.4 \cdot 10^{-3}$</td>
<td>$1.1 \cdot 10^{-3}$</td>
</tr>
</tbody>
</table>

$^1$ Estimated parameters, their relative standard deviation (%), and the residual sum of squares (RSS) over sample size $n$ were applied to nine experiments containing $n = 2786$ samples. Relative standard deviations of independent parameters were determined using the covariance matrix. Error propagation was used to determine relative standard deviations of derived parameter values. Compare supplementary material for details.

$^2$ Model 1: Michaelis-Menten, model 2: macrokinetic model without inactivation, model 3: microkinetic model without inactivation, model 4: microkinetic model including substrate-dependent inactivation

$^3$ Model 1 was applied for a maximum reaction time of 6 min, resulting in 549 samples. Parameters of model 1 for the full progress curves are listed in Table A.18.

$^4$ Values for $k_1$, $k_{-1}$, $k_3$ and $k_{-3}$ were calculated from the macrokinetic parameters according to Equations 1.14 to 1.17 including propagation of uncertainty.

$^5$ Values for $k_{catf}$, $K_{mA}$, $K_{mB}$ and $K_{eq}$ were calculated from the microkinetic parameters according to Equations 1.18 to 1.21 including propagation of uncertainty.
2.2 Kinetic modeling in the BioCatNet database system

Table 2.4: Estimated kinetic parameters and the residual sum of squares (\(RSS\)) over sample size \(n\) for the benzaldehyde lyase-catalyzed self-ligation of benzaldehyde and four kinetic models\(^6\). Footnotes and descriptions are equivalent to Table 2.3.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Model 1 (^7)</th>
<th>Model 2 (^4)</th>
<th>Model 3 (^5)</th>
<th>Model 4 (^5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k_1) (\text{mM}^{-1} \text{min}^{-1})</td>
<td>-</td>
<td>135</td>
<td>134</td>
<td>245</td>
</tr>
<tr>
<td>(k_{-1}) (\text{min}^{-1})</td>
<td>-</td>
<td>3.6</td>
<td>2.2</td>
<td>3.1 (\cdot) (10^3)</td>
</tr>
<tr>
<td>(k_3) (\text{min}^{-1})</td>
<td>(142)</td>
<td>(77)*</td>
<td>(22)*</td>
<td></td>
</tr>
<tr>
<td>(k_{-3}) (\text{mM}^{-1} \text{min}^{-1})</td>
<td>(3 (\cdot) (10^3))</td>
<td>(3 (\cdot) (10^4))</td>
<td>(37)*</td>
<td></td>
</tr>
<tr>
<td>(k_{inS}) (\text{mM}^{-1} \text{min}^{-1})</td>
<td>(55)*</td>
<td>(54)*</td>
<td>(12)*</td>
<td></td>
</tr>
<tr>
<td>(k_{catf}) (\text{min}^{-1})</td>
<td>5 (\cdot) (10^3)</td>
<td>3 (\cdot) (10^3)</td>
<td>3.1 (\cdot) (10^3)</td>
<td>4.9 (\cdot) (10^3)</td>
</tr>
<tr>
<td>(K_{mA}) (\text{mM})</td>
<td>55</td>
<td>22</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>(K_{mB}) (\text{mM})</td>
<td>(28)*</td>
<td>(131)</td>
<td>(94)*</td>
<td>(25)*</td>
</tr>
<tr>
<td>(K_{eq}) (\text{mM}^{-1})</td>
<td>22</td>
<td>23</td>
<td>23</td>
<td>32</td>
</tr>
<tr>
<td>(\text{RSS}_n) (\text{mM}^2)</td>
<td>1.4</td>
<td>15</td>
<td>15</td>
<td>2.7</td>
</tr>
</tbody>
</table>

\(^6\) Estimated parameters, their relative standard deviation (%), and the residual sum of squares (\(RSS\)) over sample size \(n\) were applied to thirteen experiments containing \(n = 374\) samples.

\(^7\) Model 1 was applied for a maximum reaction time of 6 min, resulting in 110 samples. Parameters of model 1 for the full progress curves are listed in Table A.18.
2.2 Kinetic modeling in the BioCatNet database system

Figure 2.9: Fits for the dimethoxy-benzaldehyde self-ligation and the microkinetic model with substrate-dependent enzyme inactivation (model 4, lines). Symbols as described in Figure A.36. The respective parameter estimates are given in Table 2.3 with reaction conditions according to Figure A.36 and Table A.15 (Equivalent with Figure A.44).
2.2 Kinetic modeling in the BioCatNet database system

Figure 2.10: Fits for the benzaldehyde self-ligation and the microkinetic model with substrate-dependent inactivation (model 4, lines). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15 (Equivalent with Figure A.45).
3 Discussion

The results from this thesis consist of two main parts: analyses of protein sequence networks and kinetic parameter estimation. Protein sequence networks facilitate the identification of enzyme candidates in large sequence data sets, such as potentially evolvable "hub" sequences and an approximation of a fractal network dimension as a measure of evolvability (Manuscript 1). Furthermore, communities within protein sequence networks hint at percolation in protein sequence space and an overall connectedness of homologous protein sequences (Manuscript 2). Structurally equivalent positions can be identified by standard numbering schemes, as outlined for the family of $\omega$-transaminases, and used to annotate protein sequence networks (Manuscript 3). Data on protein sequence, structure and function can be integrated in the BioCatNet database system (Manuscript 4) (Reusch, 2014; Vogel, 2015), paving the way for more reproducible data mining of enzyme kinetics (Manuscript 5). The lessons learned from kinetic parameter estimation can be used to derive an overall data management strategy for upcoming research projects on the investigation of enzymatic sequence-structure-function relationships.

3.1 From protein families to protein networks

Homologies between individual sequences can be represented as nodes in a weighted graph, i.e. a network connecting sequences with a predefined threshold of pairwise sequence identity (compare with Atkinson et al., 2009). Usually, these weighted networks are generated for sequences from a common protein family sharing fold or cofactor specificity. Mutations of proteins imply changes in nucleic acid sequences. Here, amino acid sequences were used instead, since they encode biochemical properties, thus representing a phenotype under selection pressure. The topological or statistical properties of protein sequence networks can be useful in identifying highly connected bridges between communities of a network (Manuscript 2) and in identifying connected hubs in sequence space (Manuscript 1). In contrast to the fixed branches inside a phylogenetic tree, sequence networks allow for multiple connections between two protein sequences, and seem thus more suitable to capture possible routes of protein evolution (compare with Atkinson et al., 2009).

Several concepts for evolution have been proposed, among them the principle of gradualism, which implies that natural systems evolve rather continuously, in contrast to strictly step-wise developments. Gradualism has also been coined by the quote "nature does not make jumps" that was attributed to Carl von Linné. Gradual evolution has been proposed for protein structure and function, with evolution of protein sequences following small stepwise changes by events such as mutation, recombination or duplication (Conrad, 1977). The imagination of a globally connected sequence space by Smith (1970) suggests also that protein evolution
occurs via small steps, which underlines that protein families can be represented or visualized as connected sequence networks. The topological properties found in protein sequence networks underlie an evolutionary concept similar to gradualism, where protein families and subfamilies emerge as connected communities within a network, rather than distinct phylogenetic branches.

3.1.1 Connectedness of protein sequence space

For the six protein superfamilies of $\alpha/\beta$-hydrolases (abH), $\beta$-hydroxyacid dehydrogenases/imine reductases (bHAD), cytochrome P450 monoxygenases (CYP), $\omega$-transaminases (oTA), short-chain dehydrogenases/reductases (SDR) and thiamine diphosphate-dependent decarboxylases (DC), the respective community size distributions of known sequence space have been investigated and found to follow a power law, with the extrapolated Fisher exponent $\tau_{100}$ being an upper limit to the Fisher exponent of the extant sequence space (Manuscript 2).

Since the Fisher exponent measures the ratio of small to large communities, it can be interpreted as an indicator of the global connectedness within the known sequence space of a given protein family. The general observation of few communities containing many sequences might relate with the assumption that more stable protein folds are more evolvable, thus forming larger and higher connected communities of mutations (Deeds et al., 2003).

The protein superfamilies oTA, SDR, bHAD and abH had a smaller $\tau_{100}$ (2.3, 2.4, 2.5 and 2.6, respectively) and thus a higher ratio of large to small communities than the protein superfamilies DC, or CYP (with $\tau_{100}$ of 2.8 and 3.3, respectively). A high ratio of large to small communities indicates a high connectedness within a protein sequence network. There are at least three possible reasons for a high connectedness of a protein family:

1. A high fraction of a protein superfamily’s extant sequence space has already been explored.
2. The protein superfamily shows high microdiversity, i.e. many sequences are known to be connected by point mutations.
3. The protein family covers only a small region of the total sequence space with a low overall variability in sequence homology.

The observation that the connectedness gradually increased as more sequences become known is supported by the concept of gradual saturation of sequence space. This concept describes the observation that the number of newly sequenced genes that form separate communities plotted over time decreases to zero (compare with Nelson, 2011). Rather than expanding, the sequence space of protein families seems to gradually become denser and more connected. As $\tau_{100}$ measures the connectedness of a protein family, it also measures the current level of saturation, with the investigated protein families SDR and CYP having the highest and lowest saturation of their sequence space.
3.1 From protein families to protein networks

The six protein superfamilies abH, bHAD, CYP, DC, oTA and SDR showed a similar linear dependency of $\tau$ on the clustering threshold. Thus, many small communities were observed for high threshold values, which gradually combined into larger communities with decreasing sequence identity threshold, as bridges between communities appeared. These bridging sequences were formed at decreasing sequence identity thresholds by sequences that had been part of one community and then became part of another community, or by previously isolated sequences. These bridging sequences might be interesting enzyme candidates, as they possibly share biochemical properties of both communities. If global sequence identity relates to biochemical function, a community is characterized by a similar function that differs from other communities. Bridging sequences with similarities to at least two communities are therefore promising candidates for substrate ambiguity (Jensen, 1976) or even catalytic promiscuity (Khersonsky et al., 2006).

3.1.2 Hub regions in protein sequence space

The evolution of protein sequences occurs in iterative steps of random mutagenesis of the genotype and subsequent selection of the phenotype. Therefore, the sequence space that has been iteratively explored during four billion years of evolution is expected to be connected (Smith, 1970). Since the number of explored protein sequences (estimated $10^{40}$) is much smaller than the number of theoretical sequences (estimated $> 10^{300}$), the dimension of the sub-space of extant protein sequences is expected to be much smaller than the multi-thousand dimensional space of theoretical sequences. An estimation of the dimension of the known sequence space was achieved by counting numbers of neighbors at increasing distances. The fractal network dimension $D_f$ of a protein family was similar among the investigated protein families. $D_f$ varied between 0.7 and 1.0 for sequence identities between 98 and 90%, whereas $D_f$ increased to values between 3.5 and 4.5 at lower sequence identities between 70 and 30%. The observation of a distance-dependent fractal dimension of sequence space gives an interesting insight into the sequence-function relationships of proteins. For uncorrelated random mutations, it has been estimated that the probability of protein inactivation is 34% for each mutation (Guo et al., 2004). Therefore, for a small number of mutations, the chance of finding active mutants is high ($0.66^2 = 44\%$ and $0.66^4 = 19\%$ for two and four mutations, respectively). Thus, many combinations of random mutations result in active proteins and $D_f \approx 4.0$ as evaluated for the point-mutation network of TEM $\beta$-lactamases is a lower limit of the dimension of the extant sequence space for a small number of mutations, because $D_f$ is expected to further increase as more TEM $\beta$-lactamase sequences are discovered in the future. In contrast, if 10% of all positions are randomly exchanged, the chance of finding an active variant of a 300 amino acid protein reduces to $0.66^{30} = 4 \cdot 10^{-6}$. Therefore, the mutations that result in an active protein
must be highly correlated, and evolution is dominated by the non-additive effects of epistasis (Wu et al., 2016). The high correlation of mutations is compatible with the much lower fractal network dimension $D_f$ between 0.7 and 1.0 that seems to be a generic property of all investigated protein families. For lower sequence identities between 70 and 30% the mutations become more uncoupled, which results in a considerable increase of the fractal network dimension $D_f$. At a first glance, scale-dependent network dimensions are counter-intuitive. However, scale-dependent spatial dimensions have also been observed for physical systems such as turbulent interfaces (Catrakis and Dimotakis, 1996) and for the distribution of luminous matter in the universe (Bak and Chen, 2001). Although the analysis of the distance dependence of protein sequence space is based on a relatively small number of known sequences, it provides quantitative estimates which are in agreement with known sequence-function relationships (Guo et al., 2004). It will be interesting to see how $D_f$ develops in the future when many more protein sequences become known.

Two complementary neighborhood definitions were applied to construct sequence networks. A network construction based on point mutations allowed for an interpretation of alternative evolutionary paths along the network (Zeil et al., 2016). However, mutation-based networks are restricted to the rare families with high microdiversity such as TEM β-lactamases. In contrast, the metric of global sequence identity can be applied to all protein families. For TEM β-lactamases, the mutation-based and the identity-based degree distributions were identical and approximated by a power law distribution with a scaling exponent $\gamma = 1.2$. A power law degree distribution was also observed for four protein families with low microdiversity (bHAD, DC, oTA, SDR) when using the distance metrics of pairwise sequence identity. Although the four families had different structural folds, domain arrangements, and sequence lengths, and differed in their level of sequence diversity (Figure A.3) and their size (Table 2.1), they resulted in similar scaling exponents of $\gamma$ between 1.1 and 1.3. The observation that different protein families show similar scaling exponents indicates that the constraints governing protein evolution are similar for all proteins (Keller, 2005). Scale-free distributions of protein families have been described previously for networks of co-occurring protein domains and networks of sequence motifs, with scaling exponents $\gamma$ in the range from 1.7 to 2.0 (Wuchty, 2001; Aziz et al., 2016). By clustering sequences into homologous families, scale-free cluster size distributions have been observed with scaling exponents between 1.6 and 2.5 (Koonin et al., 2002; Enright et al., 2003; Orengo and Thornton, 2005). It has been suggested that cluster size distribution is a direct consequence of the necessity for a functional protein to fold into a stable structure (Deeds et al., 2003). As a consequence, sequence space is highly connected, as seen for families with high microdiversity (Zeil et al., 2016). Connectivity is also related to findability of genotypes (McCandlish, 2013). Stability against random errors, another feature attributed to scale-free networks, is also favorable during evolution (compare with Albert and
3.1 From protein families to protein networks

Barabási, 2002).

Scale-free distributions have been found in all domains of life sciences, and far-reaching conclusions have been drawn which were not supported by the data (Lima-Mendez and van Helden, 2009). Therefore, the goodness of the power law fit was compared to alternative fits by Poisson and Gaussian distributions. While their parameters could be adjusted to follow the data in the tail, they failed to describe the monotonous increase of the number of nodes at decreasing degrees, and thus confirmed the power law fit (see also Keller, 2005; Lima-Mendez and van Helden, 2009). However, the limited number of sequences per protein family and the small fraction \(10^{-20}\) of known protein sequences (Dryden \textit{et al}., 2008) are two factors that favor the tendency to form a power law distribution, because it had been observed that binning of the data has the tendency to form a power law distribution (Jeong \textit{et al}., 2000) and that sub-networks tend to exhibit a power law distribution, irrespective of the topological property of the larger network they were sampled from (Han \textit{et al}., 2004). Even if the sub-networks retain the power law shape, their scaling exponent might differ from the larger network (Stumpf \textit{et al}., 2005). Thus, there is a risk that the apparent power law distribution might result from a sampling artifact. As the number of newly sequenced genomes is rapidly expanding in the near future, it will be interesting to see whether the degree distribution is robust upon better sampling of the sequence space.

3.1.3 Evolutionary constraints for protein sequences

By analyzing the known sequence space, it was predicted that extant proteins form a percolating, highly connected network where each sequence has multiple neighbors, and each pair of sequences is connected by many different paths, as expected from evolution (Smith, 1970). However, the density in sequence space is not uniform, but follows a power law distribution which indicates that certain folds were more evolvable than others. Percolation allows for the concept of evolution as adaptive walks on a fitness landscape (Kauffman and Levin, 1987), where sequences at the ends of the walks may substantially differ from one another (Frenkel and Trifonov, 2007). A high degree of connectedness also overcomes the possible blockage by sign epistasis and reciprocal sign epistasis (Wu \textit{et al}., 2016) and thus is a necessary condition of efficient evolution, despite the fact that only an infinitesimally small portion of the theoretical sequence space had been explored during the course of life on earth (Dryden \textit{et al}., 2008). In a highly connected sequence network as a model of evolution (Manrubia and Cuesta, 2015), sequences were found that form bridges between two communities. Since the number of bridges is much smaller than the number of community members, they only gradually appear as the number of sequenced genes increases. Consequently, the observed separation of families is merely a consequence of our limited knowledge of extant sequence space. With increasing sequence data
3.1 From protein families to protein networks

from genomics and metagenomics projects, one can expect more and more sequences to occur which form bridges between yet separated families and thus contribute to the connectedness of known sequence space. These bridging sequences are equivalent to reconstructed ancestral sequences in binary trees (Merkl and Sterner, 2016). Since they form a link between two branches, ancestral proteins are assumed to be generalists with a broader substrate spectrum or even multiple activities (Khersonsky et al., 2006). While the binary tree model of evolution assumes that the ancestor sequences have disappeared from the biosphere, the network model of evolution assumes that bridging sequences still exist. For any two neighboring, biochemically distinct communities, one expects bridging sequences to exist that contribute to the formation of a continuous network. It will be challenging to analyze how the biochemical properties change as one walks across the bridges. Most probably, bridging sequences are multi-functional or promiscuous enzymes with known or latent activities of both sub-families. In contrast to ancestors, these generalists already exist in the biosphere and are waiting to be found.

Protein networks with a highly inhomogeneous, exponential degree distribution with a long tail have another interesting consequence: the existence of a few highly connected nodes. These hubs are sequences or groups of sequences (hub regions) with a very large number of potentially functional neighbors. The role of hub regions in evolution is still under discussion. It has been suggested that highly connected nodes originated early in evolution (Fell and Wagner, 2000), while less connected nodes are recent results from divergent evolution (Dokholyan et al., 2002). This interpretation of "the old get richer" is based on preferential attachment network models (see Barabási and Albert, 1999). However, preferential attachment is only one way to generate networks, and there are different network topologies that all result in a power law degree distribution (Lima-Mendez and van Helden, 2009). As a consequence, the most highly connected protein sequences are not necessarily the phylogenetically oldest. By assuming that evolution has reached an equilibrium in protein sequence space, the more evolvable folds might have become densely populated as a consequence of convergent evolution (Dokholyan et al., 2002), thus connecting the concept of hubs to the concept of evolvability. Evolvability of a protein sequence has two aspects: robustness toward possible deleterious effects of mutations and innovability, where additional mutations readily induce new functions (Dellus-Gur et al., 2015). Since the hub sequences have many supposedly functional neighbors, they have proven to be highly evolvable. Interestingly, some hub proteins have a pivotal role in metabolism. The E1 subunit of the pyruvate dehydrogenase complex, a hub of the DC network, is also a hub in the metabolism linking glycolysis and citric acid cycle (Gray et al., 2014; Zhang et al., 2014). The aspartate aminotransferase, a hub of the oTA network, links the amino acid and the carbohydrate metabolisms (Korla et al., 2015). These coincidences of hubs in sequence networks and metabolic networks could point at a higher robustness against mutations to preserve cellular function. The concept of hubs can also be applied to improve the efficiency of
directed evolution experiments. Directed evolution is a powerful and widely applied strategy for improving biochemical and biophysical properties of proteins by applying iterative rounds of random mutations and screening. However, multiple random mutations tend to result in inactive proteins with a probability of 92% for only six random mutations (Guo et al., 2004). Therefore, it has been suggested to start a directed evolution experiment either from a population of neutral mutants (Gupta and Tawfik, 2008) or by constructing ancestor sequences (Merkl and Sterner, 2016), which are believed to have a higher robustness and thus higher evolvability than contemporary sequences (Gaucher et al., 2008). As a promising alternative, one could suggest to use the hub sequences as starting points in directed evolution experiments and to select highly evolvable homologues directly from the pool of contemporary sequences.

3.2 From data management to model management

The lessons learned from the effect of different kinetic model equations on the outcome of the parameter estimation can be generalized for the application of models on original data in general (compare with Figure A.28). The data on enzyme sequence, structure and function used in white biotechnology can be analyzed by various models (exemplified in Figure 3.1). Models refer to concepts and assumptions that are applied on original data on the level of sequence, structure or function. The application of models results in model-dependent information that cannot be understood correctly without annotations on the original data and the models themselves. The influence of different models on the outcome of the parameter estimations in Manuscript 5 emphasizes the necessity to store information on both data and models. Furthermore, other models used in data mining should be integrated in databases as well, leading to the concept of model bases instead of plain databases. Such a strategy for data and model management for white biotechnology should be implemented in subsequent steps, based on the demands of collaborative research projects combining computational routines (modeling) from fields like bioinformatics with original data from public databases or experiments.

3.2.1 Models for bioinformatic sequence analyses

Original data on sequences can be analyzed by various models, such as standard numbering schemes based on profile hidden Markov models (reviewed by Eddy, 1998). The concept of protein sequence networks from Atkinson et al. (2009) can be regarded as an exemplary application of a model: an assumption on the relationships between homologous sequences, visualized in the topology of a network. Similarly, protein families can be represented by profile HMMs, as a further model-application on the level of protein sequences (Eddy, 1998; Finn et al., 2015). As examples for models working on sequence data, protein sequence networks and standard numbering schemes have been applied in chapters 2.1.1 to 2.1.3. The combination of a stan-
3.2 From data management to model management

Figure 3.1: Concept for data and model integration in a common data- and modelbase on the levels of protein sequence (1), structure (2) and experimental time-course data (3). Models refer to concepts and assumptions that are applied on the original data, such as: alignments that are applied on protein sequences to construct protein sequence networks, standard numbering schemes (based on profile HMMs of structure-guided sequence alignments) or equations of kinetic models. The application of models results in model-dependent information (compare with Figure A.28) that cannot be reproduced and interpreted without annotations on the original data and the models themselves. Thus, routines of data mining rely on both data and model quality. Exemplary results of data mining could be the identification of hubs or bridges in protein sequence networks, conservation analyses based on standard numbering schemes or kinetic parameter estimations.

Since models like profile hidden Markov models or assumptions on sequence homology are often used for the annotation of sequences, the biological relevance of the analyzed data in sequence databases depends on both data and models. Thus, it becomes necessary to annotate both sequence data and models. As an example for the annotation of a model, the underlying standard numbering scheme (annotating structurally equivalent positions) with protein sequence networks (representing distances between sequences) helps to identify protein sequences within large data sets that match sequence motifs and are related to different subgroups of homologous sequences, as exemplified for ω-transaminases from fold type IV (Figure 2.7) (Manuscript 3).
alignments for a profile hidden Markov model of a standard numbering scheme have to be known in order the interpret the applicability of such a model. For protein sequence networks, the programs and parameters used to construct such a network have to be reported as well to avoid misinterpretation of results. A management strategy for bioinformatic models will help to reproduce data mining results, similar to the data management strategies supported by the FAIR data principles (Wilkinson et al., 2016).

3.2.2 Managing time-course data and kinetic models

The kinetic parameter estimations for PfBAL-catalyzed symmetric carboligations suggested the microkinetic model for substrate-dependent enzyme inactivation (model 4, Equations 1.13 and 1.22) as the best fit (Figures 2.9 and Figures 2.10). These observations would not have been possible with parameter estimations based on initial reaction rates, since the initial phase of the reaction could also be fitted by a two-substrate Michaelis-Menten model (model 1, Equation 1.5, Figures A.38 and A.39). The model fit and the parameter quality of the microkinetic model 4 with substrate-dependent inactivation support the mechanism of inactivation for both substrates as observed previously (Leksawasdi et al., 2004). Michaelis-Menten model 1, which had a similar fit quality and similar $k_{\text{catf}}$ values, cannot capture inactivation since it neglects the later phases of the time-courses. These findings underline the importance of full time-course data for proper investigations on enzyme kinetics (Duggleby and Morrison, 1978; Duggleby, 2001). Furthermore, it confirms the warning from Hill et al. (1977) that data from literature does not often support the Michaelis-Menten model. The better model fit of the Michaelis-Menten model 1 to the initial phases in comparison to the full time-courses does not guarantee that this model sufficiently describes the underlying mechanism (Stroberg and Schnell, 2016). When using time-courses for estimating kinetic parameters, the biochemist avoids the information loss of the rate regression from concentration over time and is rewarded by more precise information about the full enzymatic reaction, in contrast to investigations of the initial reaction phase or steady-state kinetics (Schnell and Maini, 2003).

In summary, one can propose that the selection of a suitable kinetic model should start with knowledge on the stoichiometry, which in turn determines the choice of possible reaction mechanisms (Figure 1.2) do not necessarily result in parameter estimates of the same quality. In summary, one can propose that the selection of a suitable kinetic model should start with knowledge on the stoichiometry, which in turn determines the choice of possible reaction mechanisms. Finally, different equations describing the same reaction mechanism can be necessary to estimate parameters with lower uncertainties.
3.2 From data management to model management

3.2.3 Data integration for bioinformatics and white biotechnology

As Lapatas et al. (2015) pointed out, experimenters and users have to be included in a data management strategy. Accordingly, a Microsoft Excel spreadsheet was provided on the BioCatNet documentation website for facilitated sharing of data from experimenters (Manuscript 4). To further enhance the usability of BioCatNet, standardized vocabularies could be integrated as naming conventions for terms describing enzyme function, since consistent names and annotations are required for an applicable data integration (Goble and Stevens, 2008; Wilkinson et al., 2016). Thus, ongoing projects on data integration and data mining should make use of predefined names, such as the suggestions of the Gene Ontology (GO) for various biologically relevant terms and the relations between these names (Ashburner et al., 2000). Existing guidelines, such as the descriptions of the STRENDA consortium (Tipton et al., 2014) and the proposal for reporting data on biocatalytic reactions by Gardossi et al. (2010) mentioned earlier are comparable examples of standards for data integration, which also inspired the concept of BioCatNet (Reusch, 2014; Vogel, 2015). Hence, the need for standardization on data annotation is comparable in bioinformatics and white biotechnology in general, since various data sources have to be combined for investigations on protein sequence, structure and function. The usage of common vocabularies will help to integrate different modeling routines with different data on various levels, paving the way for more reproducible data mining in data-driven modeling routines (Figure 3.1).
4 List of publications and manuscripts

Manuscripts


---

\(^\text{10}\) Manuscript 6 is not part of the main results of this work.

\(^\text{11}\) Manuscript 7 is not part of the main results of this work.
Conference posters

Presenters’ names are underlined.


Student supervision


5 Bibliography


5 Bibliography


5 Bibliography


Bibliography


5 Bibliography


Bibliography


6 Erklärung der eigenständigen Arbeit

Hiermit versichere ich, die vorliegende Arbeit selbständig und ohne Verwendung anderer als der bezeichneten Hilfsmittel und Quellen verfasst zu haben. Aus fremden Quellen entnommene Passagen sind als solche kenntlich gemacht.

Patrick C. F. Buchholz
Stuttgart, 7. Mai 2018
A Appendix: Publications

The following publications comprise additional results and details on data and computational methods for chapter 2.

- chapter A.1 (Manuscript 1) refers to chapter 2.1.1.
- chapter A.2 (Manuscript 2) refers to chapter 2.1.2.
- chapter A.3 (Manuscript 3) refers to chapter 2.1.3.
- chapter A.4 (Manuscript 4) refers to chapter 2.2.1.
- chapter A.5 (Manuscript 5) refers to chapter 2.2.2.
A.1 The scale-free nature of protein sequence space


**Own contributions**

I established the computational methods for the generation and analysis of protein sequence networks and applied them for different superfamilies (see chapter 2.1.1). I contributed to the writing of the original manuscript.

A.1.1 Abstract

The sequence space of five protein superfamilies was investigated by constructing sequence networks. The nodes represent individual sequences, and two nodes are connected by an edge if the global sequence identity of two sequences exceeds a threshold. The networks were characterized by their degree distribution (number of nodes with a given number of neighbors) and by their fractal network dimension. Although the five protein families differed in sequence length, fold, and domain arrangement, their network properties were similar. The fractal network dimension $D_f$ was distance-dependent: a high dimension for single and double mutants ($D_f = 4.0$), which dropped to $D_f = 0.7 - 1.0$ at 90\% sequence identity, and increased to $D_f = 3.5 - 4.5$ below 70\% sequence identity. The distance dependency of the network dimension is consistent with evolutionary constraints for functional proteins. While random single and double mutations often result in a functional protein, the accumulation of more than ten mutations is dominated by epistasis. The networks of the five protein families were highly inhomogeneous with few highly connected communities (“hub sequences”) and a large number of smaller and less connected communities. The degree distributions followed a power-law distribution with similar scaling exponents close to 1. Because the hub sequences have a large number of functional neighbors, they are expected to be robust toward possible deleterious effects of mutations. Because of their robustness, hub sequences have the potential of high innovability, with additional mutations readily inducing new functions. Therefore, they form hotspots of evolution and are promising candidates as starting points for directed evolution experiments in biotechnology.

A.1.2 Introduction

Power laws of the form $f(x) \sim x^\gamma$ are ubiquitous in many physical systems and describe scale free phenomena, for which changing the scale of the independent variable $x$ preserves the functional form $f$ of the solution ($f(\lambda x) = \lambda^\gamma f(x)$) (Newman, 2005). Because scaling is a manifestation of the dynamics and geometry of a physical system, scaling laws reflect underlying
A.1 The scale-free nature of protein sequence space

generic features and provide insight into important universal principles, characterized by the scaling exponent $\gamma$.

Power laws also play an important role in life sciences. Spanning many orders of magnitude, fundamental variables such as metabolic rate, growth rate, or tree height follow a power law with an exponent $\gamma$ which is an integer multiple of $\frac{1}{4}$ (West and Brown, 2004). The observation of scaling relationships throughout the living world has inspired the search for basic principles that explain complex biological phenomena from unicellular organisms to trees. Power laws also describe population genetics for unlinked loci in the monomorphic limit and are a consequence of Darwin’s theory of evolution (Manhart et al., 2012). For proteins, scaling relations were observed for the solvent-accessible surface area (Moret et al., 2009), the packing (Reuveni et al., 2008), and the equilibrium dynamics (Tang et al., 2017), and it has been suggested that near-criticality might be a characteristic of biological systems (Mora and Bialek, 2011).

Power law distributions have also been detected in sequence similarity networks of proteins (Enright et al., 2003) and have been interpreted as a consequence of evolution (Deeds et al., 2003; Koonin et al., 2002) or the constraints of protein structure (Deeds et al., 2003; Wuchty, 2001). Detailed modelling of protein evolution is challenging due to the high complexity of combining random genotypic variation with selection of phenotypic traits such as folding pathway, protein stability, and biological function of the protein. Usually, the effects of mutations are non-additive and dominated by epistasis (Wu et al., 2016). Moreover, only an infinitesimally small fraction of the sequence space of proteins has been inspected yet, despite the rapidly growing amount of DNA data due to advances in DNA sequencing techniques. While we currently know the sequences of $10^8$ proteins (Bateman et al., 2015), the number of different protein sequences existing in the biosphere was estimated to be $10^{34}$, and up to $10^{43}$ different protein sequences might have been explored during 4 Gyr of evolution (Dryden et al., 2008). Though this number seems to be large, it is extremely small as compared to the number of possible protein sequences ($10^{300}$ theoretical sequences for a medium-sized protein). Thus, we only know a tiny fraction of the total sequence space of viable proteins, and the theoretical sequence space is sparsely populated by the extant proteins.

In the absence of knowledge about the extant sequence space, relationships between known sequences can be measured by a metric based on global sequence identity or by neighborhood relationships in a protein sequence network where sequences form nodes that are connected by edges (Widmann and Pleiss, 2014; Zeil et al., 2016). While pairwise sequence identity can be determined for all protein families, extended protein sequence networks only exist for families with high microdiversity such as TEM $\beta$-lactamases, which form a single connected network of more than 260 single point variants. In this network, the number of neighbors of each sequence is not equally distributed, but follows a power law with a scaling exponent of 1.2 (Zeil et al., 2016). The protein sequence network of TEM $\beta$-lactamases contains a few ”hubs” such as
A.1 The scale-free nature of protein sequence space

TEM-1 and TEM-116 (Jacoby and Bush, 2016) and a large number of less connected nodes, with about 10 times less sequences having 10 times more neighbors each. It is tempting to relate the property of being a highly connected node to the property of being an ancestral sequence by intuitively assuming a preferential attachment model of network generation (Barabási and Albert, 1999). However, the observed scale-free degree distribution can result from a variety of different mechanisms (Lima-Mendez and van Helden, 2009) and might be determined by the actual constraints of the system rather than a unique mechanism (Fox Keller, 2005).

A central constraint in protein evolution is the evolvability of a protein sequence, which includes two elements, robustness to faults and innovability (Dellus-Gur et al., 2013). Innovability seems to be a consequence of active site location (Dellus-Gur et al., 2013). Robustness can be measured by the tolerance of a protein for deleterious effects of mutations and is related to stability and conformational dynamics of a protein (Tokuriki and Tawfik, 2009; Dellus-Gur et al., 2015). Thus, robustness is expected to vary between and inside a protein family, and it is desirable to identify or construct highly evolvable protein family members as promising starting points for directed evolution experiments.

A.1.3 Methods

Datasets of protein sequences  The datasets of the individual protein families were updated by performing BLAST searches against the non-redundant protein database from the NCBI (GenBank) (Benson et al., 2018). The sequence datasets were updated for the families of TEM β-lactamases (TEM, 422 sequences), β-hydroxyacid dehydrogenases/ imine reductases (bHAD, 30781 sequences), thiamine diphosphate-dependent decarboxylases (DC, 39290 sequences), ω-transaminases (oTA, 120921 sequences), and short-chain dehydrogenases/ reductases (SDR, 141496 sequences). In case of TEM β-lactamases, the core region from positions 24 to 280 was used only (referring to TEM-1 position numbering).

Sequence alignments and sequence networks  The distances between pairs of protein sequences can be measured either by counting point mutations or by pairwise sequence alignments. The former metric was applied for the densely connected family of TEM β-lactamases, for which a single point mutation forms the minimal distance between two sequences. TEM β-lactamase protein sequences were connected by an edge, if they differed by one point mutation, which was feasible due to the high microdiversity of this protein family.

Pairwise distances between sequences of the remaining protein superfamilies were calculated by combining the heuristic alignment approach of USEARCH, which reduced the number of sequence pairs, with global Needleman-Wunsch sequence alignment (Edgar, 2010; Needleman and Wunsch, 1970). USEARCH alignments were performed to identify highly similar neighbor sequences with an identity threshold of 0.5, corresponding to 50% sequence identity without
A.1 The scale-free nature of protein sequence space

Terminal gaps. In the second step, more accurate global sequence identities were derived from pairwise Needleman-Wunsch alignments (implemented in the EMBOSS bioinformatics software suite (Rice et al., 2000)) with gap opening penalty of 10 and gap extension penalty of 0.5. USEARCH and EMBOSS were run on multiple threads by applying GNU Parallel (Tange, 2011).

The point-mutation network of TEM β-lactamases and the identity-based networks of the remaining protein superfamilies, i.e. the sequence networks calculated by global sequence alignments as described above, were constructed and visualized by Cytoscape (version 3.4.0) using prefuse force directed layout. For the identity-based networks, prefuse force directed layout was applied with respect to the edge weights (i.e. the higher the sequence identity, the closer the sequences are placed).

**Degree distribution and fractal network dimension** For identity-based sequence networks, the number of neighboring sequences for a given sequence, i.e. the degree of a network node, was determined by counting the number of sequence pairs having a minimum sequence identity to the respective sequence, such as ≥ 95% sequence identity and thus less than 5% pairwise distance. For the point mutation network of TEM β-lactamases, the degree of a network node was determined by counting neighboring sequences within the distance of one point mutation to the respective sequence. The degrees were calculated for all sequences of a given sequence network and the number of sequences \( N \) having \( n \) neighbors was plotted over the degree \( n \).

To derive the fractal network dimension \( D_f \) of identity-based sequence networks, the number of sequence pairs \( p(d) \) that differed by less than \( d \% \), with \( d \% = (100 - \text{sequence identity}) \% \), was computed for pairwise sequence identities determined by USEARCH (Edgar, 2010). The respective fractal network dimension \( D_f \) was calculated assuming \( p(d) \sim d^{D_f} \) and plotting \( \log(p(d)) \) over \( \log(d) \) for \( d = 2, 4, 6, \ldots, 100 \). In addition, \( p(d) \) was determined for the point-mutation network of TEM β-lactamases with \( d = 1, 2, 3, 4 \) point mutations.

A.1.4 Results

All members of a protein family are related to each other by their global sequence identity obtained from pairwise sequence alignments. This relationship was analyzed by constructing networks where the nodes represent individual sequences and the edges represent a neighborhood relationship. Two types of neighborhood relationships were applied for network construction. In identity-based networks, an edge is formed between a pair of sequences if their global sequence identity exceeds a threshold. By adjusting the sequence identity threshold, the construction of identity-based sequence networks was feasible for all homologous protein families and resulted in connected networks for each family. In the rare case of protein families with high
A.1 The scale-free nature of protein sequence space

microdiversity, such as the TEM β-lactamase family, a second network type was constructed, where an edge between two nodes was formed if the two sequences differed by a point mutation. If such a point mutation network is feasible, it is expected to be highly similar to the respective identity-based with high sequence identity threshold.

Network models for TEM β-lactamases, a family of high microdiversity The TEM β-lactamase family is a large protein family of high microdiversity (Zeil et al., 2016). A point mutation network was constructed for variants of the TEM β-lactamase core region, resulting in 267 nodes and 401 edges (Figure A.1). The number of neighbors varied widely for each node. While there were two highly connected hubs (TEM-1 with 86 and TEM-116 with 55 neighbor sequences), most nodes had only few neighbors. The network properties were characterized by calculating the degree distribution, and the number $N$ of nodes with $n$ neighbors followed a power law distribution $N(n) \sim n^{-\gamma}$ with a scaling exponent $\gamma = 1.2$ (Figure A.2).

For comparison, an identity-based network was constructed using a global sequence identity threshold of 99.5% pairwise sequence identity, corresponding to a distance of one point mutation (Figure A.6). The global sequence identity measures the number of mutations between two sequences, but is independent of the number of known sequences between the two sequences. The network consisted of 267 nodes and 401 edges, too, and its degree distribution followed a power law with a scaling exponent $\gamma$ which was identical to the point mutation network (Figure A.7).

Alternatively, the degree distributions of the sequence networks were fitted by a Poisson distribution $P(\lambda)$ and a Gaussian distribution $G(\mu, \sigma)$. In contrast to the power-law distribution, the Poisson and the Gaussian distribution resulted in noticeably qualitative deviations from the experimental data (Figure A.7).

Degree distributions for protein superfamilies with low microdiversity Except for the TEM β-lactamases, the microdiversities of the protein families were too low to result in connected point mutation networks. Therefore, four protein superfamilies (β-hydroxyacid dehydrogenases/ imine reductases, bHAD; thiamine diphosphate-dependent decarboxylases, DC; ω-transaminases, oTA; short-chain dehydrogenases/ reductases, SDR) were analyzed by constructing networks based on pairwise sequence identity (Table A.1). The protein families differed in their distributions of pairwise sequence identities, which is expected for superfamilies of different sequence length, fold, and domain arrangement (Figure A.3).

Sequence pairs with a global sequence identity $\geq 95\%$ were defined as neighbors. For the four identity-based networks, the degree distribution was approximated by a power law distribution $N(n) \sim n^{-\gamma}$, whereas the distributions deviated from a scale-free behavior for the most highly connected nodes (Figure A.4). Thus, data for degrees $\geq 50$ or 70 were excluded.
from linear regression, resulting in scaling exponents of $\gamma = 1.2 - 1.3$ (Table A.2). The power law distribution was maintained upon decreasing the global sequence identity thresholds for the construction of identity-based networks to $\geq 90\%$, $\geq 85\%$, or $\geq 80\%$ (Figures A.8 to A.10), and the scaling exponents $\gamma$ decreased slightly with decreasing threshold to $\gamma = 0.9 - 1.1$. Furthermore, subsets between 10\% and 90\% randomly selected sequences from the DC superfamily resulted in similar scaling exponents $\gamma$ between 1.1 and 1.4 (Table A.4).

The inhomogeneous power law degree distributions of identity-based sequence networks point to the existence of highly connected hubs in the sequence space of the four protein superfamilies (Table A.3). Instead of individual hub sequences, communities of highly connected nodes with similar degrees were identified in the identity-based networks. For the DC superfamily, the 100 most highly connected protein sequences had between 250 and 266 neighboring sequences. Upon random selection of a subset of protein sequences from the DC superfamily, the respective sequences with the highest number of neighboring sequences were found to be highly similar, unless very small subsets were analyzed (Table A.5).

**Dimensions of protein sequence networks** As a further network property, the fractal network dimension $D_f$ was evaluated by counting the number of sequence pairs $p(d)$ that differed by less than $d\%$ (100\% - sequence identity) for $d = 2, 4, 6, \ldots$ (Figure A.5). For low values of $d$ ($d \leq 10\%$, i.e. $\geq 90\%$ identity), $\log p(d)$ increased linearly with $\log d$, resulting in a network dimension $D_f = 0.7 - 1.0$ for the four superfamilies with low microdiversity (Table A.2). Random selection of a subset of protein sequences from the DC superfamily lead to almost identical values of $D_f \approx 0.7$ for $d \leq 10\%$ (Figure A.11). For increasing distance $d$, the network dimension $D_f$ increased to $D_f = 3.5 - 4.5$ for $30\% \leq d \leq 70\%$. For the family of TEM $\beta$-lactamases, $D_f$ was estimated to 1.8 from the values at $d = 2\%$ and $d = 4\%$.

Because of the high sequence identities of the members of the TEM $\beta$-lactamase family, only few sequence pairs showed distances higher than 4\%. Estimating the fractal network dimension for the point-mutation network of TEM $\beta$-lactamases by comparing the number of single and double mutants resulted in a higher value of $D_f = 4.0$ (Figure A.12). Beyond double mutants, the limited network size resulted in an apparent decrease of the network dimension, and the analysis of double, triple, and quadruple mutants resulted in $D_f = 1.8$, as observed for the identity-based TEM $\beta$-lactamase network.

**A.1.5 Discussion**

**The dimension of protein sequence space** The evolution of protein sequences occurs in iterative steps of random mutagenesis of the genotype and subsequent selection of the phenotype. Therefore, the sequence space that has been iteratively explored during 4 Gyr of evolution is expected to be connected (Smith, 1970). Since the number of explored protein sequences
A.1 The scale-free nature of protein sequence space

(10^{40}) is much smaller than the number of theoretical sequences (> 10^{300}), the dimension of the sub-space of extant protein sequences is expected to be much smaller than the multi-thousand dimensional space of theoretical sequences. An estimation of the dimension of the known sequence space was achieved by counting the numbers of neighbors at increasing distances. The fractal network dimension $D_f$ of a protein family was similar among the investigated protein families. $D_f$ varied between 0.7 and 1.0 for sequence identities between 98 and 90 %, whereas $D_f$ increased to values between 3.5 and 4.5 at lower sequence identities between 70 and 30 %. The observation of a distance-dependent fractal dimension of sequence space gives an interesting insight into the sequence-function relationships of proteins. For uncorrelated random mutations, it has been estimated that the probability of protein inactivation is 34 % for each mutation (Guo et al., 2004). Therefore, for a small number of mutations, the chance of finding active mutants is high ($0.66^2 = 44 \%$ and $0.66^4 = 19 \%$ for two and four mutations, respectively). Thus, many combinations of random mutations result in active proteins, and $D_f \approx 4.0$ as evaluated for the point-mutation network of TEM β-lactamases is a lower limit of the dimension of the extant sequence space for a small number of mutations, because $D_f$ is expected to further increase as more TEM β-lactamase sequences are discovered in the future. In contrast, if 10 % of all positions are randomly exchanged, the chance of finding an active variant of a 300 amino acid protein reduces to $0.66^{30} = 4 \cdot 10^{-6}$. Therefore, the mutations that result in an active protein must be highly correlated, and evolution is dominated by the non-additive effects of epistasis (Wu et al., 2016). The high correlation of mutations is compatible with the much lower fractal network dimension $D_f = 0.7 - 1.0$, which seems to be a generic property of all investigated protein families. For lower sequence identities between 70 and 30 %, the mutations become more uncoupled, which results in a considerable increase of the fractal network dimension $D_f$.

At a first glance, scale-dependent network dimensions are counter-intuitive. However, scale-dependent spatial dimensions have also been observed for physical systems such as turbulent interfaces (Catrakis and Dimotakis, 1996) and for the distribution of luminous matter in the universe (Bak and Chen, 2001). Although the analysis of the distance dependence of protein sequence space is based on a relatively small number of known sequences, it provides quantitative estimates which are in agreement with known sequence-function relationships (Guo et al., 2004). It will be interesting to see how $D_f$ develops in the future, when many more protein sequences become known.

**Evolutionary constraints for protein sequence space** Two complementary neighborhood definitions were applied to construct sequence networks. A network construction based on point mutations allows for an interpretation of alternative evolutionary paths along the network (Zeil et al., 2016). However, mutation-based networks are restricted to the rare families with high microdiversity such as TEM β-lactamases. In contrast, the metric of global sequence...
identity can be applied to all protein families. For TEM $\beta$-lactamases, the mutation-based and the identity-based degree distributions were identical and were approximated by a power law distribution with a scaling exponent $\gamma = 1.2$. A power law degree distribution was also observed for four protein families with low microdiversity (bHAD, DC, oTA, SDR) when using the distance metrics. Although the four families have different structural folds, domain arrangements, and sequence lengths, and differ in their level of sequence diversity (Figure A.3) and their size (Table A.1), they resulted in similar scaling exponents $\gamma = 1.2 - 1.4$. The observation that different protein families show similar scaling exponents indicates that the constraints that govern protein evolution are similar for all proteins (Fox Keller, 2005).

Scale-free distributions of protein families have been described previously for networks of co-occurring protein domains and networks of sequence motifs, with scaling exponents $\gamma$ in the range from 1.7 to 2.0 (Wuchty, 2001; Aziz et al., 2016). By clustering sequences into homologous families, scale-free cluster size distributions have been observed with scaling exponents between 1.6 and 2.5 (Enright et al., 2003; Koonin et al., 2002; Orengo and Thornton, 2005; Buchholz et al., 2017). It has been suggested that cluster size distribution is a direct consequence of the necessity for a functional protein to fold into a stable structure (Deeds et al., 2003). As a consequence, sequence space is highly connected, as seen for families with high microdiversity (Zeil et al., 2016). Connectivity is also related to findability of genotypes (McCandlish, 2013). Stability against random errors, another feature attributed to scale-free networks, is also favorable during evolution (Albert and Barabási, 2002).

**Pitfalls and limitations for protein sequence networks** While scale-free distributions seem to be ubiquitous in many domains of life sciences, care should be taken when drawing far-reaching conclusions which are not supported by the data (Lima-Mendez and van Helden, 2009). Therefore, the goodness of the power law fit was compared to alternative fits by Poisson and Gaussian distributions. While the parameters of the Poisson and Gaussian distributions could be adjusted to follow the data in the tail, they fail to describe the monotonous increase of the number of nodes at decreasing degrees, and thus confirm the power law fit (Lima-Mendez and van Helden, 2009; Fox Keller, 2005). However, the limited number of sequences per protein family and the small fraction ($10^{-20}$) of known protein sequences (Dryden et al., 2008) are two factors that favor the tendency to form a power law distribution, because it has been observed that binning of the data has the tendency to form a power law distribution (Jeong et al., 2000) and that sub-networks tend to exhibit a power law distribution, irrespective of the topological property of the larger network they were sampled from (Han et al., 2004). By analyzing randomly selected sub-networks, we demonstrated that the scaling exponent was robust upon resampling, thus excluding the possibility that the scaling exponent might differ between network and sub-networks (Stumpf et al., 2005). However, there is still a risk that
the apparent power law distribution might result from a sampling artifact. As the number of newly sequenced genomes is rapidly expanding in the near future, it will be interesting to see whether the degree distribution is robust upon better sampling of the sequence space.

**Implications for protein evolution and protein engineering**  
Protein networks with a highly inhomogeneous, exponential degree distribution with a long tail have another interesting consequence: the existence of a few highly connected nodes. These hubs are sequences or groups of sequences with a very large number of potentially functional neighbors.

The role of hubs in evolution is still under discussion. It has been suggested that highly connected nodes originated early in evolution (Fell and Wagner, 2000), while less connected nodes are recent results from divergent evolution (Dokholyan et al., 2002). This interpretation of “the old get richer” is based on preferential attachment network models (Barabási and Albert, 1999). However, preferential attachment is only one way to generate networks, and there are different network topologies which all result in a power law degree distribution (Lima-Mendez and van Helden, 2009). As a consequence, the most highly connected protein sequences are not necessarily the phylogenetically oldest, thus hub sequences should not be interpreted as ancestors. By assuming that evolution has reached an equilibrium in protein sequence space, the more evolvable folds might have become densely populated as a consequence of convergent evolution (Dokholyan et al., 2002), thus connecting the concept of hubs to the concept of evolvability. The observation of a uniform distribution of sequences from thermophilic and hyperthermophilic sources in the oTA network demonstrated that hub sequences are not characterized by increased thermostability (Figure A.26, Buß et al., 2018)

Evolvability of a protein sequence has two aspects: robustness toward possible deleterious effects of mutations and innovability, where additional mutations readily induce new functions (Dellus-Gur et al., 2013). Since the hub sequences have many supposedly functional neighbors, they have proven to be highly evolvable. Interestingly, some hub proteins have a pivotal role in metabolism. The E1 subunit of the pyruvate dehydrogenase complex, a hub of the DC network, is also a hub in the metabolism linking glycolysis and citric acid cycle (Gray et al., 2014; Zhang et al., 2014). The aspartate aminotransferase, a hub of the oTA network, links the amino acid and the carbohydrate metabolisms (Korla et al., 2015). These coincidences of hubs in sequence networks and metabolic networks could point at a higher robustness against mutations to preserve cellular function.

The concept of hubs can also be applied to improve the efficiency of directed evolution experiments. Directed evolution is a powerful and widely applied strategy for improving biochemical and biophysical properties of proteins by applying iterative rounds of random mutations and screening. However, multiple random mutations tend to result in inactive proteins with a probability of 92% for only six random mutations (Guo et al., 2004). Therefore, it has been
A.1 The scale-free nature of protein sequence space

suggested to start a directed evolution experiment either from a population of neutral mutants (Gupta and Tawfik, 2008) or by constructing ancestor sequences (Merkl and Sterner, 2016) which are believed to have a higher robustness and thus higher evolvability than contemporary sequences (Gaucher et al., 2008). As a promising alternative, we suggest to use the hub sequences as promising starting points in directed evolution experiments and to select highly evolvable homologues directly from the pool of contemporary sequences.

Acknowledgements

We thank Jannik Seidel for his explorative studies, Silvia Fademrecht for providing the sequence family databases, and Uta Freiberg for inspiring discussions.

Figure A.1: Sequence network for 267 TEM β-lactamases formed by 401 point mutations (edges) with 259 sequences forming a densely connected network with two hub sequences (TEM-1 depicted as black rectangle, TEM-116 as black oval). First neighbors of hub sequences are depicted in dark gray, other sequences in white.
A.1 The scale-free nature of protein sequence space

Figure A.2: Distribution of the number of sequences $N$ in a network of TEM β-lactamase point mutations having $n$ first neighbors (Figure A.1). The degree distribution follows a power law with exponent $\gamma = 1.2$. 

$\gamma_{TEM} = 1.2$
Figure A.3: Distributions of pairwise global sequence identity for the protein families from Table A.1 as determined by high-scoring sequence pairs in USEARCH (Edgar, 2010).
A.1 The scale-free nature of protein sequence space

Figure A.4: Neighbor distribution for the protein families with low microdiversity from Table A.1 with neighbors defined by ≥ 95% global sequence identity. The corresponding scale-free exponents $\gamma$ were derived from linear regression for degrees ≤ 50 (bHAD, DC) or ≤ 70 (oTA, SDR) and are summarized in Table A.2.
A.1 The scale-free nature of protein sequence space

Figure A.5: Cumulated distributions of sequence pairs $p(d)$ for pairwise distances of $d$ % of the protein families TEM (open squares), DC (filled squares), bHAD (open circles), SDR (diamonds) and oTA (filled circles) from Table A.1 in subsequent distance intervals of 2 % distance $d$ (100 % - sequence identity). Linear fits are shown as red lines for distances up to 10 % identity (up to 5 % for TEM). For further distances between 70 and 30 %, an approximately linear area is depicted in red.
A.1 The scale-free nature of protein sequence space

Table A.1: Overview of the analyzed protein family networks by number of nodes (sequences) and maximal degree (number of neighbors) for a 95% sequence identity threshold, with average sequence length. The small family of TEM β-lactamases is shown as reference due to its high microdiversity with a threshold of 99.5% sequence identity (a).

<table>
<thead>
<tr>
<th>Enzyme family (abbreviation)</th>
<th>Nodes</th>
<th>Maximal degree</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEM β-lactamases (TEM)</td>
<td>267 a</td>
<td>86 a</td>
<td>250</td>
</tr>
<tr>
<td>β-hydroxyacid dehydrogenases/ imine reductases (bHAD)</td>
<td>17020</td>
<td>259</td>
<td>320</td>
</tr>
<tr>
<td>thiamine diphosphate-dependent decarboxylases (DC)</td>
<td>24880</td>
<td>266</td>
<td>580</td>
</tr>
<tr>
<td>ω-transaminases (oTA)</td>
<td>79987</td>
<td>381</td>
<td>460</td>
</tr>
<tr>
<td>short-chain dehydrogenases/ reductases (SDR)</td>
<td>81680</td>
<td>312</td>
<td>300</td>
</tr>
</tbody>
</table>

Table A.2: Overview of the analyzed protein families from Table A.1 and their derived parameters. The scale-free exponent \( \gamma \) refers to sequence identity networks constructed with pairwise identity thresholds of 95% (compare with Table A.4, 99.5% threshold for TEM β-lactamases^a). Network dimension \( D_f \) refers to the slope from Figure A.5 in different regions of pairwise sequence identity (> 90%).

<table>
<thead>
<tr>
<th>Enzyme family</th>
<th>( \gamma )</th>
<th>( D_f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEM</td>
<td>1.2 a</td>
<td>1.8</td>
</tr>
<tr>
<td>bHAD</td>
<td>1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>DC</td>
<td>1.1</td>
<td>0.7</td>
</tr>
<tr>
<td>oTA</td>
<td>1.2</td>
<td>0.9</td>
</tr>
<tr>
<td>SDR</td>
<td>1.3</td>
<td>1.0</td>
</tr>
</tbody>
</table>
### A.1 The scale-free nature of protein sequence space

Table A.3: Exemplary network hubs and their annotations from sequence networks with a threshold of 95% sequence identity (99.5% for TEM β-lactamases)\(^a\) for the protein families from Table A.1.

<table>
<thead>
<tr>
<th>Family</th>
<th>Annotation</th>
<th>Source</th>
<th>NCBI accession</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEM(^a)</td>
<td>β-lactamase TEM-1 2-hydroxy-3-oxopropionate reductase</td>
<td><em>Acinetobacter baumannii</em></td>
<td>AAP20891</td>
<td>86</td>
</tr>
<tr>
<td>bHAD</td>
<td>2-hydroxy-3-oxopropionate reductase</td>
<td><em>Proteobacteria</em></td>
<td>WP_001303675</td>
<td>259</td>
</tr>
<tr>
<td>DC</td>
<td>pyruvate dehydrogenase subunit</td>
<td><em>Gammaproteobacteria</em></td>
<td>WP_044256366</td>
<td>266</td>
</tr>
<tr>
<td>oTA</td>
<td>putrescine aminotransferase aspartate aminotransferase</td>
<td><em>Enterobacter cloacae</em></td>
<td>WP_042715413</td>
<td>381</td>
</tr>
<tr>
<td>SDR</td>
<td>GDP-mannose 4,6-dehydratase</td>
<td><em>Shigella</em></td>
<td>WP_000069444</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Helicobacter pylori</em></td>
<td>WP_058338748</td>
<td>312</td>
</tr>
</tbody>
</table>
A.1 The scale-free nature of protein sequence space

A.1.6 Supporting Information


In this supplementary material, abbreviations for protein families are used for TEM $\beta$-lactamases (TEM), as an exemplary family having high microdiversity, and $\beta$-hydroxyacid dehydrogenases/ imine reductases (bHAD), thiamine diphosphate-dependent decarboxylases (DC), $\omega$-transaminases (oTA) and short-chain dehydrogenases/ reductases (SDR).

![Figure A.6](image-url) Figure A.6: Sequence network for 267 TEM $\beta$-lactamases connected by 401 edges above a 99.5\% pairwise sequence identity threshold, in comparison to the point mutation network (Figure A.1). Hub sequences are depicted in black (TEM-1 as black rectangle, TEM-116 as black oval) with their first neighbors depicted in dark gray, other sequences in white.
A.1 The scale-free nature of protein sequence space

Figure A.7: Distribution of the number of sequences $N$ having $n$ first neighbors for the distance-based network of TEM β-lactamases (Figure A.6). The degree distribution hints at a power law distribution with exponent $\gamma = 1.2$ (A). In addition, probability density functions were fitted for a power-law distribution (line, $\gamma = 1.2$), a Gaussian distribution (dashed line, $\mu = 3.0$, $\sigma = 6.4$) and a Poisson distribution (dotted line, $\lambda = 3.0$) with residual sum of squares 0.01, 0.2 and 0.1, respectively (B - D).
A.1 The scale-free nature of protein sequence space

Figure A.8: Degree distribution for the protein families with low microdiversity from Table A.1 with neighbors defined by $\geq 90\%$ global sequence identity. Linear regression was performed for degrees $\leq 50$ (bHAD, DC) or $\leq 70$ (oTA, SDR).
Figure A.9: Degree distribution for the protein families with low microdiversity from Table A.1 with neighbors defined by $\geq 85\%$ global sequence identity. Linear regression was performed for degrees $\leq 50$ (bHAD, DC) or $\leq 70$ (oTA, SDR).
A.1 The scale-free nature of protein sequence space

Figure A.10: Degree distribution for the protein families with low microdiversity from Table A.1 with neighbors defined by ≥ 80% global sequence identity. Linear regression was performed for degrees ≤ 50 (bHAD, DC) or ≤ 70 (oTA, SDR).
A.1 The scale-free nature of protein sequence space

Figure A.11: Cumulative distributions of sequence pairs $p(d)$ for pairwise distances of $d$ % of the DC protein superfamily for the complete data set (open squares) and 10 %, 20 %, . . . , 90 % randomly selected subsets of sequences (filled squares). The areas marked in red correspond to the linear approximations from Figure A.5.
Figure A.12: Cumulative distribution of sequence pairs $p(d)$ with distance in $d$ point mutations of TEM β-lactamases (open squares). Linear fits are shown for $d = 1, 2$ (red line) and for $d = 2, 3, 4$ point mutations (blue line).
A.1 The scale-free nature of protein sequence space

Table A.4: Scaling exponents $\gamma$ for randomly selected subnetworks of the DC superfamily, with edges formed by a threshold of 95\% pairwise sequence identity. Linear regressions were performed up to a limited number of neighbors only, due to low sampling quality for higher degrees. Thus, values for $\gamma$ were determined up to a maximum degree.

<table>
<thead>
<tr>
<th>Selection [%]</th>
<th>$\gamma$</th>
<th>Maximal degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.1</td>
<td>50</td>
</tr>
<tr>
<td>90</td>
<td>1.2</td>
<td>50</td>
</tr>
<tr>
<td>80</td>
<td>1.2</td>
<td>50</td>
</tr>
<tr>
<td>70</td>
<td>1.2</td>
<td>30</td>
</tr>
<tr>
<td>60</td>
<td>1.1</td>
<td>30</td>
</tr>
<tr>
<td>50</td>
<td>1.2</td>
<td>30</td>
</tr>
<tr>
<td>40</td>
<td>1.1</td>
<td>20</td>
</tr>
<tr>
<td>30</td>
<td>1.2</td>
<td>10</td>
</tr>
<tr>
<td>20</td>
<td>1.2</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>1.4</td>
<td>10</td>
</tr>
</tbody>
</table>

Table A.5: Exemplary protein sequences found in hub regions of the DC networks for varying subsets of randomly selected sequences. The Annotations are listed as "pyruvate dihydrogenase subunit" (PDH), "glyoxylate carboligase" (GLX) or "acetolactase synthase 2 catalytic subunit" (ALS). Pairwise sequence identities towards the hub sequence of the complete network (WP_044256366) are given in the column on the right.

<table>
<thead>
<tr>
<th>Selection [%]</th>
<th>Annotation</th>
<th>Source</th>
<th>NCBI accession</th>
<th>Degree</th>
<th>Identity [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>PDH</td>
<td>Gammaproteobacteria</td>
<td>WP_044256366</td>
<td>266</td>
<td>100.0</td>
</tr>
<tr>
<td>90</td>
<td>PDH</td>
<td>Citrobacter sp. MGH 55</td>
<td>WP_043001220</td>
<td>229</td>
<td>99.8</td>
</tr>
<tr>
<td>80</td>
<td>PDH</td>
<td>Gammaproteobacteria</td>
<td>WP_044256366</td>
<td>212</td>
<td>100.0</td>
</tr>
<tr>
<td>70</td>
<td>PDH</td>
<td>Citrobacter sp. MGH 55</td>
<td>WP_043001220</td>
<td>180</td>
<td>99.8</td>
</tr>
<tr>
<td>60</td>
<td>PDH</td>
<td>Enterobacteriaceae</td>
<td>WP_000815384</td>
<td>146</td>
<td>96.2</td>
</tr>
<tr>
<td>50</td>
<td>PDH</td>
<td>Gammaproteobacteria</td>
<td>WP_044256366</td>
<td>133</td>
<td>100.0</td>
</tr>
<tr>
<td>40</td>
<td>PDH</td>
<td>Escherichia coli</td>
<td>WP_021550521</td>
<td>103</td>
<td>95.8</td>
</tr>
<tr>
<td>30</td>
<td>GLX</td>
<td>Salmonella enterica</td>
<td>WP_038390089</td>
<td>68</td>
<td>24.0</td>
</tr>
<tr>
<td>20</td>
<td>PDH</td>
<td>Citrobacter sp. MGH 55</td>
<td>WP_043001220</td>
<td>51</td>
<td>99.8</td>
</tr>
<tr>
<td>10</td>
<td>ALS</td>
<td>Escherichia albertii</td>
<td>WP_025238020</td>
<td>24</td>
<td>28.1</td>
</tr>
</tbody>
</table>

125
A.2 Percolation in protein sequence space


**Own contributions**

I applied the computational methods to calculate distributions of pairwise sequence identity and community (cluster) sizes for different protein superfamilies (see chapter 2.1.2). I contributed to the writing of the original manuscript.

A.2.1 Abstract

The currently known protein sequences are not distributed equally in sequence space, but cluster into families. Analyzing the cluster size distribution gives a glimpse of the large and unknown extant protein sequence space, which has been explored during evolution. For six protein superfamilies with different fold and function, the cluster size distributions followed a power law with slopes between 2.4 and 3.3, which represent upper limits to the cluster distribution of extant sequences. The power law distribution of cluster sizes is in accordance with percolation theory and strongly supports connectedness of extant sequence space.

Percolation of extant sequence space has three major consequences:

1. It transforms our view of sequence space as a highly connected network where each sequence has multiple neighbors, and each pair of sequences is connected by many different paths. A high degree of connectedness is a necessary condition of efficient evolution, because it overcomes the possible blockage by sign epistasis and reciprocal sign epistasis.

2. The Fisher exponent is an indicator of connectedness and saturation of sequence space of each protein superfamily.

3. All clusters are expected to be connected by extant sequences that become apparent as a higher portion of extant sequence space becomes known. Being linked to biochemically distinct homologous families, bridging sequences are promising enzyme candidates for applications in biotechnology because they are expected to have substrate ambiguity or catalytic promiscuity.

A.2.2 Introduction

Despite the rapidly growing amount of DNA data due to advances in DNA sequencing techniques, only a tiny fraction of all protein sequences existing in the biosphere has been sequenced, yet. While we currently know the sequences of almost $10^8$ proteins (Bateman et al., 2015), the
number of extant sequences was estimated to be $10^{34}$, and up to $10^{43}$ different protein sequences might have been explored during 4 Gyr of evolution (Dryden et al., 2008). Though this number seems to be large, it is infinitesimally small as compared to the theoretical sequence space ($10^{400}$ possible sequences for a medium-sized protein), and it would be highly improbable to find functional proteins by random search (Salisbury, 1969). Therefore, the Darwinian model of protein evolution based on mutation of the genotype and subsequent natural selection of the phenotype excludes the possibility of extant sequences being scattered in the theoretical sequence space, but they are expected to form a connected network, where functional sequences and mutations form the nodes and edges, respectively (Smith, 1970). In his fundamental article about the structure of sequence space, J. Maynard Smith asked the questions whether indeed all existing proteins are part of a single network with a single starting point, what fraction of the functional sequence space has been explored yet, and how large is the space of functional, but never-born proteins (Chiarabelli et al., 2006). Although the sequence space of functional proteins is unknown, we can reliably measure distances between sequences by global or local alignment methods (Henikoff and Henikoff, 1992). The currently known protein sequences are not equally distributed, but cluster into homologous families (Rappoport et al., 2012). However, due to the sparsity of the known sequence space, in most clusters even neighboring nodes differ by multiple mutations. As an exception, the TEM β-lactamase family has a very high microdiversity, and the variants form a dense single network with nodes connected by single mutations (Zeil et al., 2016).

The apparent sparsity of the known sequence space is a consequence of our limited knowledge of the extant sequences in the biosphere. Therefore, we expect that as we know more sequences, all nodes will gradually form a connected network. As an alternative explanation of sparsity, the observed separation between clusters is the consequence of ancestor sequences having become extinct during evolution (Thornton, 2004).

In this work, the known sequence space was explored by applying percolation theory to learn about the extant sequence space. Percolation theory describes the cluster distribution on a randomly populated lattice, with a parameter $p$ describing the occupancy of the lattice sites (Christensen and Moloney, 2005). For increasing values of $p$, the characteristic cluster size $s_\xi$ and the fraction $P$ of sites belonging to the largest cluster increases. As $p$ approaches the percolation threshold $p_c$, an infinite cluster appears for the first time on an infinite lattice, while on a finite-sized lattice the largest cluster percolates between the lattice boundaries. The core of percolation theory is a set of scaling relations that depend on $|p_c - p|$, such as $s_\xi \sim |p_c - p|^{-1/\sigma}$ and $P \sim (p - p_c)^\beta$ with critical exponents $\sigma$ and $\beta$ which depend on the geometry of the lattice. Most importantly, percolation theory predicts that the cluster size distribution $N(s)$ (the number $N$ of clusters with size $s$) decreases for $s \ll s_\xi$ as $N(s) \sim s^{-\tau}$ and decays exponentially for $s \gg s_\xi$. Near to percolation ($p \to p_c$), $s_\xi$ becomes infinite. Thus,
for $s$ spanning many orders of magnitude $\log N(s)$ depends linearly on $\log s$, with the Fisher exponent $\tau$ describing the ratio of small to large clusters.

Thus, investigating the cluster size distribution $N(s)$ of homologous protein families provides insights into the structure of the known sequence space and gives a glimpse of the extant sequence space, despite our limited knowledge.

### A.2.3 Materials and methods

**Clustering** The in-house databases on $\alpha/\beta$ hydrolases (abH, 395000 sequences) (Pleiss et al., 2000), cytochrome P450 monoxygenases (CYP, 53000 sequences) (Gricman et al., 2015), thiamine diphosphate-dependent decarboxylases (DC, 39000 sequences) (Vogel and Pleiss, 2014), and $\beta$-hydroxyacid dehydrogenases/imine reductases (bHAD, 31000 sequences) (Fademrecht et al., 2016) were updated by searching the NCBI non-redundant protein database (GenBank (Benson et al., 2011)) by BLAST (Camacho et al., 2009). For each homologous family, representative sequences were selected as seed sequences. Family databases for short-chain dehydrogenases/reductases (SDR, 141000 sequences) and $\omega$-transaminases (oTA, 121000 sequences) were established based on seed sequences derived from literature (Persson and Kallberg, 2013; Rudat et al., 2012). For each protein database, sequence identities of high-scoring sequence pairs were calculated by the USEARCH software suite (version 9.2) (Edgar, 2010). Sequence pairs with a distinct sequence identity cutoff were clustered by the Python module graph-tool (version 2.17)\(^{12}\).

**Cluster size distribution** For the six protein superfamilies, the cluster size distribution $N(s)$ was analyzed for cluster sizes $s=1, 2, 3, \ldots, 1000$. Because for large cluster sizes data becomes increasingly sparse, a histogram distribution was generated by counting the number of clusters $N_{i,j} = \sum_{s=i}^{j} N(s)$ with cluster sizes between $i$ and $j$.

The observed cluster size distributions were compared to three model distributions: a Gaussian distribution $N(s) \sim \exp \left(\frac{1}{2} \cdot (s - \mu)^2 / \sigma^2 \right)$, an exponential distribution $N(s) \sim \exp (-b \cdot s)$ and a power law distribution $N(s) \sim s^{-\tau}$ with the Fisher exponent $\tau$ characterizing the model distribution (Figure A.16). Excel sheets for the calculation of the distributions are provided as supporting information\(^{13}\). The log-log plots of the three model distributions differ considerably: $\log N(s)$ of the Gaussian distribution increases gradually with $\log s$ and decays rapidly for $s > \mu$, while for the exponential distribution is decays rapidly for all $s > 0$. In contrast, for the power law distribution $\log N(s)$ depends linearly on $\log s$ with a slope of $-\tau$.

For each model distribution, the respective histogram distribution was calculated. Qualitatively, the histogram distributions were similar to the model distributions. For power law


\(^{13}\) https://doi.org/10.1371/journal.pone.0189646.s006, accessed on April 30, 2018.
A.2 Percolation in protein sequence space

distributions with $\tau > 1$, the corresponding histogram distribution could also be approximated by a straight line with a slope of $-\tau_h$. However, the two slopes $-\tau$ and $-\tau_h$ deviated.

For each histogram distribution of the six protein families, the slope $-\tau_h$ was determined by fitting the initial linear decay ($N_{1,10}$, $N_{11,100}$, and $N_{101,1000}$) by linear regression, and the Fisher exponent of the respective cluster size distribution was derived from $\tau_h$ by varying $\tau$ of the model distribution to fit the observed $\tau_h$.

A.2.4 Results

Sequence space The known protein sequence space is rapidly increasing, but it represents only a tiny fraction of the extant sequence space, that has been explored during evolution. In turn, the extant sequence space represents a fraction $p$ of the much bigger sequence space coding for functional proteins. Although both the extant and the functional sequence space and therefore also $p$ are unknown, the scaling properties of the cluster size distribution can be used as an indicator of $p$: if the cluster size distribution in the extant sequence space follows a power law over many orders of magnitude, $p$ is close to a critical percolation threshold $p_c$.

The scaling properties of the extant sequence space are investigated by analyzing the scaling properties of the much smaller space of known sequences. Because a typical protein superfamily consists of $10^4 - 10^5$ protein sequences, the cluster size range is limited to 2 - 3 orders of magnitude. The sparsity of the known sequence space has three major consequences:

1. Because of the poor statistics of the cluster size distribution $N(s)$ between $s = 1$ and 1000, the number of clusters with a size between 1 and 10 ($N_{1,10}$), 11 and 100 ($N_{11,100}$), and 101 and 1000 ($N_{101,1000}$) are analyzed, and the corresponding cluster size distribution is derived from this histogram distribution.

2. Except for very few families, e.g. TEM $\beta$-lactamases, it is rare that two members of a protein superfamily differ by only one amino acid. Therefore, neighbor relationships are established by global sequence identity as a cutoff criterion. Using a 90% cutoff criterion, two proteins of 400 amino acids are considered to be neighbors if they differ in less than 40 positions. As a consequence, the structure of the resulting network and the Fisher exponent $\tau$ depend on the cutoff criterion for the neighborhood relationship.

3. The Fisher exponent $\tau$ depends on the number of known sequences. As the number of known sequences increases, the protein families become more densely populated, and the number of large clusters is expected to increase. As a consequence, the Fisher exponent $\tau$ decreases. Therefore, the observed Fisher exponent $\tau$ as evaluated from the known protein superfamilies represents an upper limit to the Fisher exponent of the extant sequence space.
A.2 Percolation in protein sequence space

The structure of the known sequence space was analyzed for six large protein superfamilies with high diversity in sequence and function: α/β hydrolases (abH, 395,000 sequences) (Pleiss et al., 2000), short-chain dehydrogenases/reductases (SDR, 141,000 sequences), ω-transaminases (oTA, 121,000 sequences), cytochrome P450 monooxygenases (CYP, 53,000 sequences) (Gricman et al., 2015), thiamine diphosphate-dependent decarboxylases (DC, 39,000 sequences) (Vogel and Pleiss, 2014), and β-hydroxyacid dehydrogenases/imine reductases (bHAD, 31,000 sequences) (Fademrecht et al., 2016) (Table A.6). The six protein superfamilies differ in their fold and their number of family members, which is reflected in the distributions of pairwise sequence identity (Figure A.13). In the abH superfamily, the majority of sequences had pairwise sequence identity of 40 - 60 %, while almost all CYPs had a pairwise sequence identity of 15-25 %. SDRs, DCs and bHADs showed a bimodal distribution with maxima at 20 - 30 and 40 - 50 %.

Cluster size distribution For each of the six protein superfamilies, the sequences were clustered by a cutoff criterion of 60 % global sequence identity which is often applied for defining homologous families. The number $N$ of clusters with size $s$ was analyzed in a histogram with logarithmic bins for $s$ between 1 and 10, 11 and 100, 101 and 1,000, 1,001 and 10,000, and 10,001 and 100,000 to improve statistical sampling (Figure A.14). Intuitively, we had expected a Gaussian normal distribution, assuming a random distribution of cluster sizes. However, in contrast to intuition, the distribution of cluster sizes followed a power law $N(s) \sim s^{-\tau_h}$, indicated by a linear dependency of log $s$ and log $N(s)$ for the six protein superfamilies (abH, SDR, oTA, CYP, DC, bHAD). The Fisher exponent $\tau_h$ of a histogram describes the ratio between small and large clusters and is derived from linear regression in the log-log plot of the histogram (Fisher, 1967). From the Fisher exponent $\tau_h$ of the histogram, the Fisher exponent $\tau$ of the underlying cluster size distribution was calculated by fitting the observed $\tau_h$ of the histogram to a model distribution of cluster sizes following a power law distribution. Though the protein families differ in size, structure, and function, for four of the five (SDR, oTA, DC, bHAD) the Fisher exponent $\tau$ varied only slightly (1.8 - 1.9). The smallest Fisher exponent was derived for the CYP superfamily ($\tau = 1.6$). For the largest superfamily (abH), the Fisher exponent was 2.0. These values are in agreement with the Fisher exponent of $\tau \approx 2$ determined for the protein family size distribution of the Gene3D database (Orengo and Thornton, 2005) or the TRIBES resource (Enright et al., 2003), while the distribution of protein folds showed a slightly larger exponent of 2.5 (Koonin et al., 2002).

Dependency of $\tau$ on the cluster criterion While the Fisher exponent $\tau$ was almost independent of the protein family and its size, its absolute value depended on the cutoff criterion used for clustering. Upon clustering of the six families with six cutoffs between 60 and 90 %,
the cluster size distributions followed a power law for all cutoffs (Figure A.17). With increasing clustering cutoff, the relative number of small clusters increases, while the number of large clusters decreases. Consequently, the Fisher exponent $\tau$ increased almost linearly with increasing cutoff (Figure A.15) from $\tau_{60} = 1.6 - 2.0$ at 60% cutoff, to $\tau_{90} = 2.2 - 2.9$ at 90% cutoff. The Fisher exponent $\tau$ was extrapolated to a cutoff of 100%, representing a network of nodes separated by only one mutation ($\tau_{100}$). For the six protein families, the extrapolated $\tau_{100}$ values varied between 2.4 and 3.3 (2.6, 2.4, 2.3, 3.3, 2.8 and 2.5 for abH, SDR, oTA, CYP, DC, and bHAD, respectively).

Dependency of $\tau$ on the number of sequences Of the 395,000 abH sequences, 50, 25, or 12.5% were randomly selected and clustered, and the cluster size distribution was determined for four distinct cutoff values (Figure A.18). With a decreasing number of sequences, the relative number of small clusters increased, while the number of large clusters decreased. Consequently, the Fisher exponent $\tau$ increased with decreasing number of sequences: at 60% cutoff from 2.0 for the complete database to 2.0, 2.2, and 2.3 at 50, 25, and 12.5% randomly selected abH sequences, respectively. A similar trend was observed for the other cutoff values: $\tau_{70} = 2.1 - 2.5$, $\tau_{80} = 2.2 - 2.9$, $\tau_{90} = 2.4 - 3.2$. Therefore, it is expected that the Fisher exponent $\tau$ of the cluster distribution of the known sequences decreases as more extant sequences will be sequenced in the future, and the extrapolated $\tau_{100}$ values for the six families (between 2.4 and 3.3) represent upper limits to the cluster size distribution of the extant sequence network. Because percolation theory predicts values of $\tau$ between 2.055 for percolation in a 2-dimensional lattice and 2.5 in a lattice with more than 5 dimensions (Saberi, 2015), the upper limits of 2.4 - 3.3 are in agreement with percolation in the extant sequence space.

Thus, the observation of a power law cluster size distribution results from the connectedness of extant sequence space which is as a consequence of Darwinian evolution. Interestingly, a model that describes protein structural evolution on a three dimensional lattice also results in a power law cluster size distribution with an exponent of 2.3 (Deeds et al., 2003). It is a tempting observation that the two foundations of protein evolution, the connectedness of extant sequence space and the formation of a stable fold, both result in a power law cluster size distribution with a similar exponent. This observation relates to the fundamental property of protein folds: the stability of a fold is closely related to its evolvability. The more stable a fold is, the more sequences can adopt it, thus forming larger and better connected sequence networks.

A.2.5 Discussion

Connectedness and saturation of sequence space The cluster size distribution of the known sequence space of six protein superfamilies followed a power law, with the extrapolated
A.2 Percolation in protein sequence space

Fisher exponent $\tau_{100}$ being an upper limit to the Fisher exponent of the extant sequence space. The observation of few clusters containing many sequences might relate with the assumption that more stable protein folds are more evolvable, thus forming larger and higher connected clusters of mutations. The extrapolated Fisher exponent is independent of characteristic properties of the protein families such as family size (Table A.6). Because the Fisher exponent measures the ratio of small to large clusters, it can be interpreted as an indicator of the global connectedness of the known sequences of a protein family. The protein families oTA, SDR, bHAD and abH ($\tau_{100} = 2.3, 2.4, 2.5$ and $2.6$, respectively) had a smaller $\tau_{100}$ and thus a higher ratio of large to small clusters than the protein families DC, or CYP ($\tau_{100} = 2.8$ and $3.3$, respectively). A high ratio of large to small clusters indicates a high connectedness. There are at least three possible reasons for a high connectedness of a protein family:

1. The protein family is well explored; thus, a high fraction of its extant sequence space is already known.

2. The protein family has a high microdiversity.

3. The protein family covers only a small region in sequence space, thus overall variability is low.

Our observation that the connectedness gradually increased as more sequences become known is supported by the concept of gradual saturation of sequence space. This concept describes the observation that the number of newly sequenced genes that form separate clusters plotted over time decreases to zero (Nelson, 2011). Rather than expanding, the sequence space of protein families is gradually becoming denser and more connected. As $\tau_{100}$ measures the connectedness of the protein family, it also measures the current level of saturation, with the protein families SDR and CYP having the highest and lowest saturation, respectively.

Bridges between homologous families The six protein families showed a similar linear dependency of $\tau$ on the clustering cutoff. Thus, for high cutoff values many small clusters were observed, which gradually combine into larger clusters as the clustering cutoff was decreased, and bridges between clusters gradually appeared (Figures A.19 and A.20). These bridges were formed by sequences that had been part of one cluster and then became part of a second cluster, or were recruited from previously isolated sequences, as the clustering cutoff was decreased. These bridging sequences are interesting, as they belong to both clusters. If global sequence similarity relates to biochemical function, a cluster is characterized by a similar biochemical function that differs between the clusters. The bridging sequences, having similarities to two or even more clusters, are therefore promising candidates with substrate ambiguity (Jensen, 1976) or even catalytic promiscuity (Khersonsky et al., 2006).
A.2 Percolation in protein sequence space

**Protein evolution**  By analyzing the known sequence space, we predict that extant proteins form a percolating, highly connected network where each sequence has multiple neighbors, and each pair of sequences is connected by many different paths, as expected from evolution (Smith, 1970). However, the density in sequence space is not uniform, but follows a power law distribution which indicates that certain folds were more evolvable than others. Percolation allows for the concept of evolution as adaptive walks on a fitness landscape (Kauffman and Levin, 1987), where sequences at the ends of the walks may substantially differ from one another (Frenkel and Trifonov, 2007). A high degree of connectedness also overcomes the possible blockage by sign epistasis and reciprocal sign epistasis (Wu et al., 2016) and thus is a necessary condition of efficient evolution, despite the fact that only an infinitesimally small portion of the theoretical sequence space been explored during the course of life on Earth (Dryden et al., 2008). In a highly connected sequence network as a model of evolution (Manrubia and Cuesta, 2015), sequences are found that form bridges between two clusters. Since the number of bridges is much smaller than the number of cluster members, they only gradually appear as the number of sequenced genes increases. Consequently, the observed separation of families is merely a consequence of our limited knowledge of extant sequence space. With increasing sequence data from genomics and metagenomics projects, we expect more and more sequences to occur which form bridges between yet separated families and thus contribute to the connectedness of known sequence space.

These bridging sequences are equivalent to reconstructed ancestral sequences in binary trees (Merkl and Sterner, 2016). Since they form a link between two branches, ancestral proteins are assumed to be generalists with a broader substrate spectrum or even multiple activities (Khersonsky et al., 2006). While the binary tree model of evolution assumes that the ancestor sequences have disappeared from the biosphere, the network model of evolution assumes that bridging sequences still exist. For any two neighboring, biochemically distinct clusters, we expect bridging sequences to exist that contribute to the formation of a continuous network. It will be challenging to analyze how the biochemical properties change as we walk across the bridges. Most probably, bridging sequences are multi-functional or promiscuous enzymes with known or latent activities of both sub-families. In contrast to ancestors, these generalists already exist in the biosphere and are waiting to be found.

**Acknowledgements**

The authors acknowledge Lenz Lorenz (University of Stuttgart) for providing the cluster analysis tool, Uta Freiberg (University of Stuttgart) for inspiring discussions, and the German Research Foundation DFG (FOR1296, EXC310) for financial support.
A.2 Percolation in protein sequence space

Figure A.13: Distributions of pairwise global sequence identity for the protein families of $\alpha/\beta$-hydrolases (abH), short-chain dehydrogenases/ reductases (SDR), $\omega$-transaminases (oTA), cytochrome P450 monooxygenases (CYP), thiamine diphosphate-dependent decarboxylases (DC) and $\beta$-hydroxyacid dehydrogenases/ imine reductases (bHAD).
A.2 Percolation in protein sequence space

Figure A.14: Cluster size distribution of α/β hydrolases (abH), short-chain dehydrogenases/reductases (SDR), ω-transaminases (oTA), cytochrome P450 monooxygenases (CYP), thiamine diphosphate-dependent decarboxylases (DC), and β-hydroxyacid dehydrogenases/imine reductases (bHAD) follow a power law distribution: \( N(s) \sim s^{-\gamma} \) (\( N(s) \), number of clusters of size \( s \); \( \gamma \), Fisher exponent). Cluster criterion: 60% global sequence identity.
A.2 Percolation in protein sequence space

Figure A.15: Fisher exponent $\tau$ of the size distribution of homologous families for clustering cutoffs between 60 and 90\% with extrapolated Fisher exponent $\tau_{100}$ determined by linear regression (abbreviations according to Figure A.14).

Figure A.15: Fisher exponent $\tau$ of the size distribution of homologous families for clustering cutoffs between 60 and 90\% with extrapolated Fisher exponent $\tau_{100}$ determined by linear regression (abbreviations according to Figure A.14).
Table A.6: Protein superfamily size and the Fisher exponent extrapolated to 100% sequence identity ($\tau_{100}$) of the six protein families.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Enzyme superfamily</th>
<th>Superfamily size</th>
<th>$\tau_{100}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>abH</td>
<td>$\alpha/\beta$ hydrolases</td>
<td>395000</td>
<td>2.6</td>
</tr>
<tr>
<td>SDR</td>
<td>short-chain dehydrogenases/reductases</td>
<td>141000</td>
<td>2.4</td>
</tr>
<tr>
<td>oTA</td>
<td>$\omega$-transaminases</td>
<td>121000</td>
<td>2.3</td>
</tr>
<tr>
<td>CYP</td>
<td>cytochrome P450 monooxygenases</td>
<td>53000</td>
<td>3.3</td>
</tr>
<tr>
<td>DC</td>
<td>thiamine diphosphate-dependent decarboxylases</td>
<td>39000</td>
<td>2.8</td>
</tr>
<tr>
<td>bHAD</td>
<td>$\beta$-hydroxyacid dehydrogenases/ imine reductases</td>
<td>31000</td>
<td>2.5</td>
</tr>
</tbody>
</table>
A.2.6 Supporting Information


Figure A.16: Model distributions displayed as log-log plot: Gaussian distribution $N(s) = a \cdot \exp \left( -\frac{1}{2} \frac{(s-\mu)^2}{\sigma^2} \right)$ with $a = 10000$, $\mu = 200$, $\sigma = 50$, exponential distribution $N(s) = a \cdot \exp (-bs)$ with $a = 10000$ and $b = 0.2$, power law distribution $N(s) = a \cdot s^{-\tau}$ with $a = 10000$, $-\tau = 2.5$. 

138
A.2 Percolation in protein sequence space

Figure A.17: Cluster size distributions for 60, 70, 80, and 90% global sequence identity of the six protein superfamilies from Table A.6 (α/β-hydrolases in blue, shortchain dehydrogenases/reductases in green, ω-transaminases in black, cytochrome P450 monooxygenases in red, thiamine diphosphate-dependent decarboxylases in cyan and β-hydroxyacid dehydrogenases/imine reductases in orange).
A.2 Percolation in protein sequence space

Figure A.18: Cluster size distributions for 60, 70, 80, and 90% global sequence identity of all abH sequences (filled squares) and randomly selected abH sequences: 50% (open squares), 25% (filled circles) and 12.5% (open circles) of the original dataset.
A.2 Percolation in protein sequence space

Figure A.19: Details of sequence identity networks for two homologous families of short-chain dehydrogenases/reductases (SDR) with clustering cutoff at 39% sequence identity. The network shows bridges connecting the two homologous families (indicated in red hexagons). Visualization in Cytoscape (version 3.2.1) using organic layout.

Figure A.20: Details of sequence identity networks for two homologous families of short-chain dehydrogenases/reductases (SDR) with clustering cutoff at 40% sequence identity. The bridge sequences from Figure A.19 are indicated in red hexagons. Visualization in Cytoscape (version 3.2.1) using organic layout.
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space


**Own contributions**

I established the standard numbering schemes for Fold types I and IV $\omega$-transaminases and applied them on protein sequence networks (see chapter 2.1.3). I supervised all bioinformatic analyses including the contribution of Maike Gräff who curated the oTAED. I contributed to the writing of the original manuscript.

A.3.1 Abstract

The $\omega$-Transaminase Engineering Database (oTAED) was established as a publicly accessible resource on sequences and structures of the biotechnologically relevant $\omega$-transaminases ($\omega$-TAs) from Fold types I and IV. The oTAED integrates sequence and structure data, provides a classification based on fold type and sequence similarity, and applies a standard numbering scheme to identify equivalent positions in homologous proteins. The oTAED includes 67,210 proteins (114,655 sequences) which are divided into 169 homologous families based on global sequence similarity. The 44 and 39 highly conserved positions which were identified in Fold type I and IV, respectively, include the known catalytic residues and a large fraction of glycines and prolines in loop regions, which might have a role in protein folding and stability. However, for most of the conserved positions the function is still unknown. Literature information on positions that mediate substrate specificity and stereoselectivity was systematically examined. The standard numbering schemes revealed that many positions which have been described in different enzymes are structurally equivalent. For some positions, multiple functional roles have been suggested based on experimental data in different enzymes. The proposed standard numbering schemes for Fold type I and IV $\omega$-TAs assist with analysis of literature data, facilitate annotation of $\omega$-TAs, support prediction of promising mutation sites, and enable navigation in $\omega$-TA sequence space. Thus, it is a useful tool for enzyme engineering and the selection of novel $\omega$-TA candidates with desired biochemical properties.

A.3.2 Introduction

The ubiquity of amino groups in natural products leads to a great diversity of different transaminases (TAs, E.C. 2.6.1). TAs catalyze the transfer of an amino group from an amine donor
to a carbonyl acceptor with pyridoxal-5-phosphate (PLP) as a cofactor, which is bound via its aldehyde moiety to a lysine side-chain in the active site of the enzyme (Jansonius, 1998). Because the structure of the donor and the acceptor might differ and because the reaction is reversible, the enzymes accept 2 different amines as amino donors, which is known as dual substrate recognition (Hirotsu et al., 2005). As a further consequence of the reversibility of the transamination, subsequent reaction steps or product separation are required in biosynthetic applications (Han and Shin, 2014). Other members of the family of PLP-dependent enzymes include lyases, oxidoreductases, and hydrolases (Percudani and Peracchi, 2009). Because TAs differ by their amine donor and acceptor scope, substrate specificity was used to assign TAs to 2 major families. α-transaminases (α-TAs) catalyze transfer of the amino group exclusively to a carbonyl group in α-position to a carboxyl group. The acceptor can be molecules like oxaloacetate or pyruvate, which are converted to the corresponding amino donor L-glutamate or L-α-alanine. ω-transaminases (ω-TAs) lack the selectivity toward α-keto substrates and have a wide spectrum of acceptor or donor molecules, for example, α-alanine, 1-phenylethylamine, putrescine, and other amines. α-TAs are involved in amino acid biosynthesis, which makes them interesting enzymes for the production of D- or L-amino acids. Their limitation to α-carbonyl substrates is a disadvantage for broader application of these enzymes (Cristen and Metzler, 1985; Rozzell, 1984; Breslow et al., 1988).

Applications of ω-transaminases In contrast, ω-TAs lack the dependence on a carboxyl group in the acceptor substrate and are therefore promising enzymes for the synthesis of a broad range of optically pure amines (Rudat et al., 2012; Koszelewski et al., 2010; Malik et al., 2012; Ingram et al., 2007; Tauber et al., 2013; Mathew et al., 2016c, 2015; Crismaru et al., 2013) such as the pharmaceuticals imigabalin and sitagliptin (Midelfort et al., 2013; Savile et al., 2010) rivastigmine, a-aminosteroids, mexiletine, cathine, and 3-amino-8-aza-bicyclo[3.2.1]oct-8-yl-phenylmethanone (Fuchs et al., 2012; Richter et al., 2014; Koszelewski et al., 2009; Sehl et al., 2013; Weiß et al., 2016). However, considerable enzyme engineering efforts were needed to adapt enzymes to the desired substrates. In amino acid synthesis, a special focus lies on the synthesis of optically pure β-amino acids like β-phenylalanine which can be produced using lipase-ω-TA cascade reactions or racemic resolution reactions using Fold type I-(S)-selective ω-TAs (Mathew et al., 2015; Dold et al., 2016a). Non-standard amino acids might be applied in synthetic peptides or for the synthesis of the antitumor drug paclitaxel (Dold et al., 2016a; Arvidsson et al., 2001; Alvin et al., 2014). In asymmetric synthesis, the reaction equilibrium is pushed toward the desired product by removing the co-product, by high concentrations of the amine donor or by regenerating the co-substrate. For example, in the lactate dehydrogenase - glucose dehydrogenase system, the co-product pyruvate is enzymatically removed from the reaction (Weiß et al., 2016; Pavlidis et al., 2016; Hönhe et al., 2010). By using isopropyl
A.3 The \(\omega\)-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

amine as the amine donor, the resulting acetone can be easily removed by lowering the pressure (Park et al., 2013). If ortho-xylylenediamine is used as the amine donor, it polymerizes after deamination thereby shifting the equilibrium (Green et al., 2014). Beside the reaction equilibrium, solvent stability of \(\omega\)-TAs is an important factor. Many relevant substrates exhibit low solubility in aqueous solvents. While the co-solvent DMSO is compatible with \(\omega\)-TAs, only few \(\omega\)-TAs are known to be active in organic solvents such as tert-butyl ether (Savile et al., 2010; Mutti and Kroutil, 2012). Further applications are presented in the reviews of Guo and Berglund (2016), Slabu et al. (2017) and Dold et al. (2016b). More particularly, Slabu et al. (2017) presented several approaches for synthesis of pharmaceuticals using \(\omega\)-TAs.

The most widely used \(\omega\)-TAs are from the microorganisms \textit{Vibrio fluvialis}, \textit{Chromobacterium violaceum} and \textit{Arthrobacter citreus} Dold et al. (2016b). To overcome the limitations of wild type \(\omega\)-TAs, screening, data mining, and enzyme engineering are promising strategies to develop enzymes with high stability, broad substrate specificity, and high selectivity. While literature mining is an obvious starting point for most engineering projects, its success is limited by the lack of naming conventions and position numbering of \(\omega\)-TAs, which makes it difficult to compare mutagenesis studies or conserved sequence motifs in different \(\omega\)-TAs.

**Enzyme engineering of \(\omega\)-transaminases** The substrate binding sites of \(\omega\)-TAs consist of the large O- and the small P-pocket, (Crismaru et al., 2013) also called A and B pocket, respectively (Höhne et al., 2010). Promising mutation sites of \(\omega\)-TAs were reviewed recently (Guo and Berglund, 2016). In recent years, the aim of engineering efforts is obtaining enzymes with high activity toward bulky substrates lacking a carboxyl group (Pavlidis et al., 2016). In the case of \textit{Vibrio fluvialis} \(\omega\)-TA, the activity of the enzyme toward \(\beta\)-keto-methylester was substantially increased by only 8 point mutations (Midelfort et al., 2013). The most beneficial single mutation was W57F (located in the P-pocket). Pavlidis et al. (2016) could increase activity of an \((S)\)-selective \(\omega\)-transaminase from \textit{Ruegeria} sp. TM1040 by 8900 fold. The best variant was created by introducing only 4 mutations, Y59W, Y89F, Y152F, and T231A.

**Classification of \(\omega\)-transaminases** Three classification schemes for TAs based on biochemical function, sequence, or structure have been proposed. The functional classification divided the protein sequences into \(\alpha\)-TAs and \(\omega\)-TAs (Braunstein, 1973). The sequence classification assigned TAs to 5 different aminotransferase classes based on sequence similarity (Jensen and Gu, 1996; Ouzounis and Sander, 1993; Mehta et al., 1993; Lyskowski et al., 2014; Grishin et al., 1995) or 6 Pfam groups based on profile hidden Markov models (Rausch et al., 2013; Finn et al., 2014). Evolutionary analysis led to a phylogenetic tree with \(\alpha\)-TAs in Pfam group I and \(\omega\)-TAs in Pfam group III (Rudat et al., 2012; Arvidsson et al., 2001). Based on structure, PLP-dependent enzymes have been divided into 7 different fold types, (Percudani and Peracchi, 2018).
A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

2009; Grishin et al., 1995; Schneider et al., 2000) with α- and ω- TAs found in Fold types I and IV (Grishin et al., 1995; Schneider et al., 2000; Eliot and Kirsch, 2004; Pavkov-Keller et al., 2016). Fold type I (S)-selective ω-TAs have an α-β-α-structure. In contrast, Fold type IV (R)-selective ω-TAs consist of 2 domains, a 2-layer β-sandwich and an α-β-barrel. TAs of both fold types form at least homodimers, and the active sites are located at the homodimer interface. Because both α- and ω-TAs are found in each of the 2-fold types, regioselectivity is not strictly linked to global protein structure. The substrate binding sites of Fold types I and IV are mirror images with the catalytic lysine located at the si- and re-face of PLP, respectively, resulting in the observed enantiocomplementarity of the 2-folds with Fold type I (S)-selective ω-TAs converting (S)-amines and (S)-amino acids, whereas Fold type IV (R)-selective ω-TAs convert (R)-amines and (R)-amino acids as well as branched-chain L-amino acids (Lyskowski et al., 2014; Pavkov-Keller et al., 2016; Mugford et al., 2008). The 3 classification schemes are used in parallel. TAs are named by their stereopreference ((R)/(S)-selective), their substrate specificity, their regioselectivity (α/ω-TAs), or their Pfam group (aminotransferase class), resulting in enzyme names such as: ω-aminotransferase, (S)-selective aminotransferase, aromatic amino acid TA, γ-aminobutyrate TA, ω-amino acid:pyruvate TAs, and class III aminotransferase (Midelfort et al., 2013; Shin et al., 2003; Han et al., 2015; Hwang et al., 2008; Clark et al., 2009; Steffen-Munsberg et al., 2013). A public online database on ω-TAs will be a helpful tool, connecting information about mutation sites, structure data and substrate scope, thus allowing researchers the mining of uncharacterized ω-TAs for the desired enzyme functions and predicting interesting mutation sites. The first public database for ω-TA screening was the B6-database for the description and classification of vitamin B6 dependent enzymes (Per-cudani and Peracchi, 2009). Höhne et al. (2010) and Pavlidis et al. (2016) demonstrated data mining as a successful strategy for in silico screening of (R)-selective TAs accepting bulky substrates (Weiβ et al., 2016; Pavlidis et al., 2016; Pavkov-Keller et al., 2016). To support enzyme engineering and to facilitate navigation and annotation, we established a publicly available database on ω-TAs which includes sequence information and structural data. Additionally, standard numbering schemes were established for both fold types to identify equivalent positions in homologous proteins and to compare the effects of corresponding mutations in different proteins.

A.3.3 Methods

Setup and clustering of the ω-transaminase engineering database (oTAED) The sequences of 9 representative ω-transaminases (ω-TAs) were used as query sequences against the NCBI non-redundant protein database with an E-value threshold of 10−10 (Table A.11) (Benson et al., 2011). The setup of the oTAED and the clustering of the sequences was per-
A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

formed within the BioCatNet database system (Buchholz et al., 2016). A sequence identity threshold of 98% was applied to assign sequences to proteins and a threshold of 40% sequence similarity was used to form homologous families. If the majority of sequence entries from a homologous family showed sequence lengths longer than 350 amino acids, the homologous family was assigned to the Fold type I superfamily. Consequently, homologous families were assigned to the Fold type IV superfamily if most of their sequences were shorter than 350 amino acids.

All sequences within a superfamily that could not be assigned to a homologous family were collected in a separate group. Sequences shorter than 200 and longer than 550 amino acids were discarded. If available, crystal structures from the PDB repository were assigned to the sequence entries. Multiple sequence alignments and phylogenetic trees were generated using Clustal Omega (Sievers and Higgins, 2014) and can be downloaded from a WWW-accessible user interface. The oTAED is available at 14.

Standard numbering schemes Standard numbering schemes were established for the 2 superfamilies Fold type I and Fold type IV as described previously (Vogel et al., 2012). For the respective superfamily, reference structures containing the cofactor PLP and covering the most abundant homologous families were selected (Table A.12). For Fold type IV, a structural alignment of 6 reference structures was created using STAMP (Russell and Barton, 1992). For Fold type I, 14 characterized ω-TAs were selected. The N-terminal region (positions 1 to 64 of PDB entry 2YKU), which was not resolved completely in all Fold type I reference structures, was discarded to improve the robustness of the alignment. The multiple sequence alignment of the Fold type I reference sequences from Clustal Omega was refined manually, guided by structural superimposition (Sievers and Higgins, 2014). From the manually optimized alignments of Fold type I and IV reference structures, profile hidden Markov models were created by HMMER (version 3) (Eddy, 1998). By aligning all sequences to their respective sequence profile, position numbers were transferred to Fold type I sequences from PDB entry 2YKU (ω-TA from Mesorhizobium sp. LUK) and to Fold type IV sequences from PDB entry 4CE5 (ω-TA from Aspergillus terreus).

Conservation analysis The 2 standard numbering schemes for Fold type I and Fold type IV sequences were used separately to analyze the amino acid composition of both superfamilies. A position was considered to be conserved in a superfamily, if a single amino acid was present in > 70% of all sequences.

Sequence networks Sequence networks were generated for distances between pairs of homologous sequences. To reduce the number of sequences for pairwise alignments, sequences were

A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

clustered by 30% identity using the algorithm of USEARCH (Edgar, 2010). Pairwise global sequence alignments were calculated using the implementation of the Needleman-Wunsch algorithm in the EMBOSS software suite (Rice et al., 2000). Sequence networks were created by using a cutoff of 50% pairwise sequence similarity, and the resulting networks were visualized by Cytoscape version 3.4.0 using the prefuse force-directed layout algorithm with respect to the edge weights.

A.3.4 Results

The ω-transaminase engineering database (oTAED) The ω-Transaminase Engineering Database (oTAED) consists of 2 superfamilies, Fold types I and IV. The ω-TA sequences were assigned to the 2 superfamilies by sequence length. Fold type I proteins have an average length of 432 amino acids, whereas Fold type IV proteins have an average length of 297 amino acids (Figure A.25). The oTAED includes 67,210 proteins (114,655 sequences) which were separated into 169 homologous families (HFams) based on global sequence similarity.

Fold type I The Fold type I superfamily consists of 101,738 protein sequences (89% of the oTAED entries), which were assigned to 124 homologous families. Most of the putative Fold type I (S)-selective ω-TAs belong to 1 large homologous family (HFam 239) comprising 99,559 sequences (98% of all Fold type I sequences) and 164 structures. This homologous family contains the previously characterized (S)-selective ω-TAs from Mesorhizobium sp. LUK and Variovorax paradoxus (PDB entries 4AO4 and 4AOA), which are active toward aromatic β-amino acids (Crismaru et al., 2013; Wybenga et al., 2012). Their host organisms are soil bacteria living in symbiosis with plants for fixation of nitrogen (Kaneko et al., 2000; Satola et al., 2013). A further member of HFam 239 is the ω-TA from Rugeria sp. TM1040 (PDB entry 3FCR), which was characterized as an (S)-selective ω-TA with activity toward small amino acids, α-methylbenzylamine, bicyclic acceptor molecules such as exo-3-amino-8-aza-bicyclo[3.2.1]oct-8-yl-phenylmethanone and succinic semialdehyde (Weiß et al., 2016; Steffen-Munsberg et al., 2013). Other members of HFam 239 are the ω-TA from Chromobacterium violaceum (PDB entry 4AH3), which exhibited (S)-selectivity and a broad substrate range toward amines and amino acids, as well as aldehydes and ketones as acceptor molecules (Kaulmann et al., 2007), and an ω-TA from Pseudomonas aeruginosa (UniProt entry V6A7F6) with activity towards mono- and diamines. This newly characterized enzyme converts cadaverine and spermidine and catalyzes the transfer of the amino group to aromatic ketone acceptor molecules (Galman et al., 2017). These examples demonstrate the large diversity of substrate specificities in the largest homologous family of the Fold type I superfamily. The other homologous families consist of < 400 sequences each (HFam 35: 397 sequences, HFam 134: 339 sequences and 8 structures). The respective proteins are often annotated as (S)-selective ω-TAs, mostly γ-aminobutyrate
A.3 The \( \omega \)-transaminase engineering database (\( \omega \)TAED): a navigation tool in protein sequence and structure space

aminotransferases from eukaryotic organisms.

**Fold type IV**  The Fold type IV superfamily consists of 12,917 protein sequences assigned to 45 homologous families (11 % of the \( \omega \)TAED entries). It contains sequences annotated as \((R)\)-selective \( \omega \)-TAs as well as \((R)\)-selective \( D \)-\( \alpha \)-TAs (DATA). This class of enzymes is selective for \( D \)-\( \alpha \)-amino acids like \( D \)-alanine or \( D \)-glutamate (Tanizawa et al., 1989). Furthermore, it contains \( L \)-branched-chain aminotransferases (\( L \)-BCAT) with activity toward aliphatic \( \alpha \)-amino acids like valine, leucine, and isoleucine (Yvon et al., 2000; Hutson, 2001). \( L \)-BCATs show a different enantiopreference in comparison to DATA and \((R)\)-selective \( \omega \)-TAs, presumably caused by the reverse arrangement of the substrate in the active site (Pavlidis et al., 2016). However, the annotation in public databases such as NCBI or Uniprot is often restricted to DATA or \( L \)-BCAT (Höhne et al., 2010), lacking differentiation between DATA, \( L \)-BCAT, and \((R)\)-selective \( \omega \)-TAs. The largest homologous family (HFam 11) includes 11,689 sequences (90 % of all Fold type IV sequences) and 23 annotated structures such as a branched-chain-amino-acid TA (PDB entry 4WHX) and an amino lyase with activity toward 4-amino-4-deoxychorismate (PDB entry 2Y4R) (O’Rourke et al., 2011). The second largest homologous family (HFam 10) contains 511 sequences and 9 structures. This HFam contains mostly \( \omega \)-TAs annotated as \((R)\)-selective TAs, such as the \( \omega \)-TA from *Arthrobacter* sp. (PDB entries 5FR9 and 3WWH) which was adapted by large site directed mutagenesis experiments for activity toward bulky substrates such as aromatic \( \beta \)-fluoroamines or sitagliptin, and 2 \((R)\)-selective \( \omega \)-TAs from the fungi *Aspergillus fumigatus* and *A. terreus* (PDB entries 4CHI and 4CE5) showing activity toward aromatic amines (Savile et al., 2010; Höhne et al., 2010; Lyskowski et al., 2014; Cuetos et al., 2016; Thomsen et al., 2014). The other homologous families of Fold type IV contain < 200 sequences each.

**Conserved positions**  Evolutionary conserved positions often point to structurally or functionally relevant residues. Conserved positions for the 2 fold types were determined by applying the numbering schemes, as equivalent residues have the same standard number. Though some conserved positions have already been mentioned in literature, for most of the conserved positions their biochemical role is still unknown, thus being promising targets for mutational studies in the future. By applying the standard numbering schemes for Fold types I, 44 conserved positions (present in > 70 % of the sequences) were identified (Table A.7) with position numbers according to the \( \omega \)-TA from *Mesorhizobium* sp. LUK (PDB entry 2YKU). The highly conserved positions D189 and K216 were found in all sequences of Fold type I. The side-chain of D189 is fixed by interacting with H109 and is involved in binding the cofactor PLP by a hydrogen bond between the carboxylic group and the pyridine nitrogen, and is thus essential for all PLP-dependent enzymes. The conserved K216 forms a Schiff base with the intermediate
A.3 The \( \omega \)-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

or with the cofactor PLP. The role of the highly conserved D210 in (\( S \))-selective \( \omega \)-TAs is still unknown, but it might participate in a conserved salt bridge between an \( \alpha \)-helix and a \( \beta \)-strand, since in most structures an arginine or asparagine residue is found in close distance to D210 (Figure A.21). Besides the functionally relevant positions, it is remarkable that 13 of the 44 conserved residues are glycines and 4 are prolines, most of them localized in loops (12 glycines and 2 prolines, respectively) and might therefore be involved in protein folding. For 23 conserved positions, their function is still unknown. Thirty-nine conserved positions were identified in Fold type IV (Table A.8) with position numbers according to the \( \omega \)-TA from Aspergillus terreus (PDB accession 4CE5). The highly conserved position R79 is present in all sequences and is part of the conserved PLP-binding cup formed by E213, L235, I238, and T239 (Lyskowski et al., 2014). E213 forms a hydrogen bond to the cofactor PLP and might form a salt bridge to R169 in some Fold type IV proteins (Lyskowski et al., 2014). Surprisingly, the catalytically active lysine is only conserved in 88\% of all sequences at Fold type IV standard position 180. As in Fold type I, R79 might also form a salt bridge to the conserved E61. A further salt bridge could be formed between D268 and the partly conserved K/R223. All of the 11 conserved glycines were located in loops. For 20 conserved positions, their function is still unknown.

Selectivity- and specificity-determining positions Considering that (\( S \))-selective and (\( R \))-selective \( \omega \)-TAs have a different fold, different substrate specificities, and different conserved amino acids in their substrate binding pockets, it is surprising that only 1 mutation can switch enantiopreference. By engineering the \( \omega \)-TA of Arthrobacter citreus (Fold type I), it was shown that a mutation at Fold type I standard position 328 from valine to alanine changes the enantiopreference from (\( S \)) to (\( R \)) for the substrate 4-fluoro-phenylacetone (Svedendahl et al., 2010).

Other relevant positions in Fold type I are Y108 near the cofactor PLP, W26 inside the small binding pocket, and F53.1. Position F53.1 is missing in some Fold type I proteins, but was shown to have an influence on steric hindrance of bulky substrates for the \( \omega \)-TA from Chromobacterium violaceum (Denesyuk et al., 2003). With respect to the binding mechanism of substrate and cofactor of (\( R \))- and (\( S \))-selective \( \omega \)-TAs, the mechanism of binding the phosphate group of PLP via a hydrogen bond network in the phosphate-binding cup is common to both Fold types (Figure A.22) (Lyskowski et al., 2014; Denesyuk et al., 2003; Humble et al., 2012). The planar cofactor PLP is sandwiched between Y108 and V191 in Fold type I (\( S \))-selective \( \omega \)-TAs and between L235 and F217 in Fold type IV (\( R \))-selective \( \omega \)-TAs (Midelfort et al., 2013; Sayer et al., 2014). To explore which positions are involved in substrate specificity and stereoselectivity of Fold type I (\( S \))-selective \( \chi \)-TAs, 2 sequence motifs at 17 sites and 36 Fold type I standard positions were examined, which have been described in literature to be involved in
The \( \omega \)-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

substrate interactions for different \( \omega \)-TAs (Table A.9). The standard numbering of Fold type I revealed that many positions that have been described in different enzymes are structurally equivalent. One example is Fold type I standard position 192, which was described in *Vibrio fluvialis* and *Pseudomonas putida* as positions 259 and 262, respectively. This position is also part of the P-pocket (small pocket) motif, and mutations at the mentioned positions to less bulky residues can allow larger substrate molecules inside the small substrate binding pocket (Midelfort *et al.*, 2013; Park *et al.*, 2012). It is noteworthy that multiple functional roles have been suggested in literature for the same position. Fold type I standard position 53.1 is mentioned 4 times in literature for 3 different \( \omega \)-TAs. Mutation of this position resulted in changing substrate specificity or inversion of enantiopreference. Fold type I standard position 26 was mentioned 7 times for 6 different \( \omega \)-TAs. The mutation from a large residue at this position to a smaller hydrophobic residue allows the conversion of larger aromatic and hydrophobic substrates. The previously described flipping R346 was described as an important site for dual substrate recognition (Hirotu *et al.*, 2005; Steffen-Munsberg *et al.*, 2013). It was also mentioned in a motif for the recognition of \( \alpha \)-carboxyl binding of amino-acceptor substrate and described for the \( \omega \)-TA from *Vibrio fluvialis*, *Pseudomonas* sp. strain AAC and *Silicibacter pomeroyi* (Steffen-Munsberg *et al.*, 2013; Wilding *et al.*, 2016). The flipping arginine is also known for the \( \beta \)-phenylalanine converting \( \omega \)-TAs from *Variovorax paradoxus* and *Mesorhizobium* sp. LUK. For the \( \omega \)-TA from *Sphaerobacter thermophilus*, which transfers an amino group to the \( \gamma \)-position (Mathew *et al.*, 2016c), this position is a leucine, but arginines are next to this position at 346.8 and 346.13. In contrast Mathew *et al.* (2016c) determined the position 0.41 (R41) as flipping arginine in *Sphaerobacter thermophilus*, which is located at the substrate binding pocket and not in the outer shell of the enzyme (Mathew *et al.*, 2016c; Crismaru *et al.*, 2013). Furthermore, Fold type I standard position 108 seems to be an important site for the coordination of PLP. This site was mentioned in literature 4 times (Table A.9). In comparison to Fold type I, literature information about mutations in (R)-selective \( \omega \)-TAs in the Fold type IV superfamily is sparse, and only 17 Fold type IV standard positions were described (Table A.10). Most mutation data were generated by engineering of *Arthrobacter* \( \omega \)-TA for activity against prositagliptin. Fold type IV standard position 62 is part of the small pocket. In \( \omega \)-TA from *Arthrobacter* sp. 117, the mutation of V62G increased the small pocket (Savile *et al.*, 2010). In \( \omega \)-TA from *Nectria haematococca*, V62 was described as part of a motif which mediates specificity toward (R)-amines (Sayer *et al.*, 2014). Besides that, a mutation at Fold type IV standard position 62 is proposed to increase activity toward aromatic ketone substrates (Savile *et al.*, 2010).
A.3 The \(\omega\)-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

A.3.5 Discussion

The \(\omega\)-Transaminase Engineering Database (oTAED) was implemented as a public database for navigating the sequence space of the biotechnologically relevant \(\omega\)-TAs from Fold types I and IV. Besides the oTAED, databases have been published for Fold type IV (Höhne et al., 2010) and Fold type I proteins (Steffen-Münsberg et al., 2015). The conserved positions in Fold type I and IV were analyzed by standard numbering schemes in the oTAED. For each Fold type, a standard numbering scheme allowed for the unambiguous comparison of structurally equivalent positions in different \(\omega\)-TAs described in literature. The annotation of previously identified sequence motifs and the comparison of functionally relevant positions are expected to facilitate the annotation of yet uncharacterized \(\omega\)-TAs (Rudat et al., 2012; Koszelewski et al., 2010; Guo and Berglund, 2016). The inhomogeneous size distribution of homologous families, with few large and many small families has been observed previously for several other protein superfamilies, which can be interpreted as a general property of the evolved sequence space (Buchholz et al., 2017).

Comparison of fold types I and IV

The putative \((S)\)- and \((R)\)-selective \(\omega\)-TAs of Fold type I and IV, respectively, have different sequence lengths, different folds, and lack global sequence similarity, and thus are evolutionarily separate. Despite their different folds, the substrate binding sites of both fold types consist of a large O- and a small P-pocket (Crismaru et al., 2013; Höhne et al., 2010) and the catalytically important residues are located in the same spatial arrangement with a highly conserved catalytic lysine (standard positions 216 or 180 in Fold types I or IV, respectively) pointing to the cofactor PLP from the \(si\)- or \(re\)-face (Figure A.23). Thus, in respect to the cofactor and the catalytic lysine, both active sites are mirror images to each other, which explains the observed enantiocomplementarity of Fold type I \((S)\)-selective and Fold type IV \((R)\)-selective \(\omega\)-TAs (Green et al., 2014; Grishin et al., 1995). In contrast to Fold type I \((S)\)-selective \(\omega\)-TAs, Fold type IV \((R)\)-selective \(\omega\)-TAs have no reported activity toward \(\beta\)- or \(\gamma\)-amino acids (Höhne et al., 2010; Guo and Berglund, 2016; Cuetos et al., 2016; Mutti et al., 2011; Wu et al., 2017). Enzymes from both superfamilies show activity toward the small-sized \(\alpha\)-alanine (Höhne et al., 2010; Guo and Berglund, 2016; Lyskowski et al., 2014). However, isopropyl amine (IPA) is also a small amino donor, but it is not accepted by a variety of \(\omega\)-TAs (Park et al., 2013; Iwasaki et al., 2012). Furthermore, it is striking that the PLP binding cup of Fold type IV is more strictly conserved than in Fold type I, which may be explained by the smaller number of currently known sequences for this superfamily. By integrating and annotating sequence data on \(\omega\)-TAs, the oTAED could be used as a training dataset for machine learning to characterize sequences without experimental background, which was demonstrated to be a promising method for prediction of E.C. numbers.
A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

(Li et al., 2016). Fold type I and Fold type IV enzymes are still a relatively uncharted territory with only a small fraction of characterized enzymes. By using the conservation analysis as a guideline, new functionally relevant positions might be identified and explored by protein engineering.

**Fold type I** For putative Fold type I (S)-selective ω-TAs, the standard numbering allowed the determination of functional amino acid residues within different target sequences. Residues were predicted to mediate dual substrate recognition, such as the flipping arginine at Fold type I standard positions 346 or 346.1 (Hirotsu et al., 2005) which is conserved in 11,243 sequences within Fold type I and might serve as an indicator of ω-TA activity. α-TAs are absent in the Fold type I superfamily, because sequence similarities between α-TA and (S)-selective ω-TAs are very low (Sugio et al., 1995). However, they are functionally related, since a single mutation was sufficient to change an α-TA to an ω-TA (Deszcz et al., 2015). Moreover, many enzyme families like GABA-transaminases or β-phenylalanine amine transaminases were described previously within this group (Steffen-Munsberg et al., 2015). The known fingerprints of sequence positions are helpful to predict activity toward naturally occurring or technically applied amines (Steffen-Munsberg et al., 2015). According to this fingerprint-based annotations, the largest group of ω-TAs could be classified as glutamate-1-semialdehyde-aminomutases (12,667 sequences) and the second largest as β-phenylalanine transaminases (1,527 sequences). In contrast, only 883 ω-TAs were predicted to have catalytic activity toward bulky or unusual amines, which is expectable if those amines occur seldom in nature (Deszcz et al., 2015). Surprisingly, the large group of glutamate-1-semialdehyde-aminomutases comprises aminomutases and ω-TAs, which is a misleading annotation, but not unexpected when the sequence similarity is analyzed (Steffen-Munsberg et al., 2015).

**Fold type IV** Correspondingly, Fold type IV comprises different enzyme families with high global sequence similarity, but different substrate specificity: ω-TAs, 4-amino-4-deoxychorismate lyases, D-amino acid TAs (DATA), and L-branched-chain amino acid TAs (L-BCAT), which have been identified by specific sequence motifs (Table A.13), but could not be distinguished based on global sequence similarity (Höhne et al., 2010; Steffen-Munsberg et al., 2015). Interestingly, L-BCAT, 4-amino-4-deoxychorismate lyases, and DATA enzymes are related to (R)-selective ω-TAs within Fold type IV (Höhne et al., 2010) and appear as a separate branch in the oTAED (Figure A.24). The different substrate specificities are reflected by specific binding sites. In (S)-selective L-BCATs, the α-carboxyl group is bound in the small P-pocket, while in (R)-selective TAs it is bound in the larger O-pocket (Höhne et al., 2010). As a consequence, L-BCATs show an opposite enantiopreference in comparison to DATA and (R)-selective ω-TAs, and were successfully engineered into ω-TAs accepting large aliphatic substrates (Höhne...
A.3 The \( \omega \)-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

et al., 2010; Skalden et al., 2015). Furthermore, it is surprising that the arginine residue relevant for substrate recognition is not conserved within Fold type IV. This substrate recognition site was assigned to Fold type IV standard position 128 (Skalden et al., 2015). Therefore, the Fold type IV superfamily of the oTAED probably includes further enzyme classes other than \( \omega \)-TAs or several inactive enzymes. The sequence similarity network indicates that sequences with matching positions for potential (\( R \))-selectivity did not form distinct groups, whereas sequences matching motifs for ADCL, DATA, \( \text{l-BCAT} \), or ATA activity were found in different subgroups. It is, however, difficult to predict overlapping substrate scopes between different enzyme classes. Thus, enzymes annotated as \( \text{l-BCATs} \) in public repositories might behave as Fold-type IV (\( R \))-selective \( \omega \)-TAs when exposed to non-a-amino acid substrates (Boyko et al., 2016).

**Understanding the substrate specificity of \( \omega \)-TAs** Substrate specificity depends on distinct amino acid residues in the substrate pockets (Steffen-Munsberg et al., 2013). Among them, Fold type I standard position 26 seems to be pivotal in mediating the P-pocket size. This position is relevant in \( \omega \)-TAs from *Chromobacterium violaceum*, *Ochrobactrum anthropi*, *Vibrio fluvialis*, *Pseudomonas putida*, *Bacillus megaterium*, and *Caulobacter crescentus* mentioned in at least 6 publications (Table A.9). The mutation from tryptophan to a smaller hydrophobic amino acid residue at Fold type I standard position 26 allowed for conversion of larger aromatic and hydrophobic substrates. Additionally, Fold type I standard position 53.1 is crucial for the conversion of large substrates, but this position is missing in some \( \omega \)-TAs like in the *V. paradoxus* \( \omega \)-TA (Humble et al., 2012). An exchange from a large residue like W/F to F/V/A opened the small binding pocket (P) toward larger residues and led to inverted enantioselectivity and reduced activity toward 1-phenylethyl-amine, an amine donor which is accepted by all \( \omega \)-TAs (Guo and Berglund, 2016; Schätzle et al., 2009). However, the sequence region between Fold type I standard positions 44 and 81, which is probably involved in substrate recognition, showed low sequence conservation and thus could not be reliably aligned. Fold type I standard position 108 is a promising hotspot which mediates substrate recognition. This site can be exchanged with many different amino acids (M/S/N/F/A) with varying effects on substrate specificity. It was expected that a smaller residue at position 108 would allow for higher flexibility of the PLP cofactor at the active site and decrease steric hindrance for bulky substrates (Weiß et al., 2016). It was even shown that this site has an influence on \( \alpha \)- versus \( \omega \)-TA activity by structural comparison of an \( \alpha \)- with an \( \omega \)-TA from *C. violaceum* (Deszcz et al., 2015).

**Thermostability of \( \omega \)-TAs** Besides substrate specificity, enzyme stability is a target of enzyme engineering, which can be examined using the oTAED. Robustness of an enzyme toward harsh process conditions is often linked to its thermostability, which is therefore of major interest.
A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

in enzyme design.

Furthermore, psychrophilic enzymes are interesting because of their high activity at low temperatures (Struvay and Feller, 2012). Until now, no psychrophilic ω-TA and only a few thermostable ω-TAs are known for Fold type I (Mathew et al., 2016a). Recently, 3 thermostable ω-TA genes from hot spring sources were found and characterized (Uniprot entries A0A1U9WZ51, A0A1U9WZ50, and A0A1U9WZ53). Further examples are ω-TAs from Thermomicrobiurn roseum and from Sphaerobacter thermophiles (Mathew et al., 2016c,a). The taxonomic sources of sequence entries in the oTAED were searched for matching entries in the BacDive database (release 27.02.2017) which comprises environmental conditions of the 2 domains Bacteria and Archaea (Söhngen et al., 2016). For Fold type I, 2,923 sequences from thermophilic, 1,171 sequences from hyperthermophilic, and 2,434 sequences from psychrophilic source organisms were identified. For Fold type IV, 449 sequences from thermophilic and 40 sequences from hyperthermophilic source organisms were identified. In contrast, the motifs (V/I)xLDxR and PFG (K/H)YL from Stekhanova et al. (2017) for thermostable ω-TAs matched with only 12 sequences. Sequences from extremophilic source organisms did not form separate clusters, but were distributed across the respective sequence network (Figure A.26). Noteworthy, the representative node matching the motifs from Stekhanova et al. (2017) is not necessarily surrounded by matches from thermo- or hyperthermophilic sources

Acknowledgments

The authors thank Silvia Fademrecht for having established the first version of the oTAED. This work was funded by the German Federal Ministry of Education and Research (Molecular Interaction Engineering, 031A095B) and the Deutsche Forschungsgemeinschaft (FOR1296).
A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Figure A.21: Salt bridge in ω-TA Fold type I. Showed for 4AOA, 2YKU, 5GFH, and 3NUI. The conserved D210 seems to be an important salt bridge starting point, but the corresponding salt bridge partner residue is not conserved, the distance between arginine/ asparagine and aspartate is < 3.5 Å.
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Figure A.22: The active sites of ($R$)- and ($S$)-selective $\omega$-TAs (Fold type IV and I, respectively) as viewed from the $re$- and $si$-face, respectively. The functional residues were defined according to Lyskowski et al. (2014) and to Humble et al. (2012). For ($S$)-selective- $\omega$-TA, the amino acid residues of the phosphate binding cup at position 82 are serine or threonine (Rice et al., 2000; Denesyuk et al., 2003)
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Figure A.23: Structure comparison of Fold type I and IV. The pattern of Fold type I consists of an $\alpha/\beta/\alpha$ pattern with the active site located at the interface of the homodimer (only monomer is shown). Fold type IV consists of 2 clear separated domains. Domain 1 is a 2 layer sandwich. Domain 2 consists of an $\alpha$-barrel. The active site is also located at the interface of the homodimer.
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Figure A.24: Protein sequence networks of Fold type IV sequences with (A) annotated sequences matching positions of likely ($R$)-selective transaminases depicted in black and with (B) motifs for ADCL (blue), DATA (black), L-BCAT (red), and ATA (green) from 27 and Table A.13. Nodes correspond to representative sequences of clusters formed by 30% identity in USEARCH. A cutoff of 50% pairwise sequence similarity is used to select the edges. The network is shown as force-directed layout, with pairs with higher similarity arranged in closer proximity.
A.3 The \( \omega \)-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Table A.7: Conserved positions in putative Fold type I (\( S \))-selective \( \omega \)-TA sequences with standard numbering according to the \( \omega \)-TA from *Mesorhizobium* sp. LUK (PDB accession 2YKU) and their location inside the protein structure or annotated function. Positions listed here are conserved to at least 70%.

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Conserved amino acids</th>
<th>Location/function</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>G (88 %), N (5 %)</td>
<td>loop</td>
</tr>
<tr>
<td>12</td>
<td>D (81 %), L (6 %)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>G (91 %)</td>
<td>( \beta )-turn (Crismaru et al., 2013)</td>
</tr>
<tr>
<td>20</td>
<td>D (97 %)</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>G (98 %)</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>H (75 %), Y (18 %)</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>A (83 %)</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>Q (75 %), A (10 %)</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>G (98 %)</td>
<td>backbone hydrogen bond to PLP (Crismaru et al., 2013)</td>
</tr>
<tr>
<td>83</td>
<td>E (72 %), D (10 %), V (6 %)</td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>A (83 %), S (10 %)</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>A (85 %)</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>K (70 %), R (26 %)</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>A (80 %), V (8 %)</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>H (98 %)</td>
<td>interaction with D189 (Martin et al., 2007)</td>
</tr>
<tr>
<td>110</td>
<td>G (99 %)</td>
<td>loop</td>
</tr>
<tr>
<td>151</td>
<td>A (79 %), C (9 %)</td>
<td></td>
</tr>
<tr>
<td>152</td>
<td>A (78 %), C (10 %), G (9 %)</td>
<td></td>
</tr>
<tr>
<td>156</td>
<td>E (97 %)</td>
<td>hydrogen bond to PLP (Markova et al., 2005)</td>
</tr>
<tr>
<td>157</td>
<td>P (79 %), T (7 %), A (5 %)</td>
<td>loop</td>
</tr>
<tr>
<td>160</td>
<td>G (82 %)</td>
<td>loop</td>
</tr>
<tr>
<td>163</td>
<td>G (92 %)</td>
<td>loop</td>
</tr>
<tr>
<td>185</td>
<td>L (75 %), V (13 %)</td>
<td></td>
</tr>
<tr>
<td>186</td>
<td>L (71 %), F (15 %), M (5 %)</td>
<td></td>
</tr>
<tr>
<td>187</td>
<td>I (75 %), V (20 %)</td>
<td></td>
</tr>
</tbody>
</table>
A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Standard position | Conserved amino acids | Location/function
---|---|---
189 | D (100 %) | hydrogen bond to PLP (Crismaru et al., 2013; Steffen-Munsberg et al., 2015)
190 | E (96 %) | loop
194 | G (93 %) | loop
195 | G (71 %), R (15 %), A (8 %) | loop
196 | R (82 %), V (8 %) | loop
198 | G (84 %), L (6 %) | loop
201 | A (70 %), G (15 %), S (6 %) | loop
209 | P (93 %), A (6 %) | loop
210 | D (99 %) | salt bridge to arginine (Figure A.21)
216 | K (100 %) | catalytic lysine (Crismaru et al., 2013)
221 | G (91 %) | loop
223 | P (88 %), T (5 %) | loop
250 | T (94 %), S (5 %) | backbone hydrogen bond to PLP (Crismaru et al., 2013)
253 | G (83 %), A (11 %) | loop
255 | P (79 %), A (6 %) | loop
259 | A (77 %), V (5 %) | loop
288 | L (74 %), I (7 %), F (6 %) | loop
302 | R (77 %), N (6 %) | loop
305 | G (94 %) | loop

Table A.8: Conserved positions in putative Fold IV ω-TA sequences with standard numbering according to the ω-TA from Aspergillus terreus (PDB accession 4CE5) annotated as in Table A.7.

Standard position | Conserved amino acids | Location/function
---|---|---
36 | G (86 %) | loop
44 | A (83 %) | loop
56 | G (77 %), A (17 %), S (5 %) | loop
61 | E (89 %), D (9 %) | salt bridge to standard position 79
68 | G (82 %), T (5 %) | loop

©2018 Wiley Periodicals, Inc.
### A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Conserved amino acids</th>
<th>Location/function</th>
</tr>
</thead>
<tbody>
<tr>
<td>76</td>
<td>H (98 %)</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>R (100 %)</td>
<td>PLP binding cup (Lyskowski et al., 2014)</td>
</tr>
<tr>
<td>80</td>
<td>L (84 %), F (11 %)</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>S (82 %), G (11 %)</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>N (73 %), S (9 %)</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>G (95 %)</td>
<td>loop</td>
</tr>
<tr>
<td>158</td>
<td>G (86 %)</td>
<td>loop</td>
</tr>
<tr>
<td>180</td>
<td>K (88 %)</td>
<td>catalytic lysine (Lyskowski et al., 2014)</td>
</tr>
<tr>
<td>194</td>
<td>A (83 %)</td>
<td></td>
</tr>
<tr>
<td>198</td>
<td>G (83 %)</td>
<td>loop</td>
</tr>
<tr>
<td>201</td>
<td>E (72 %), D (18 %)</td>
<td></td>
</tr>
<tr>
<td>209</td>
<td>G (89 %)</td>
<td>loop</td>
</tr>
<tr>
<td>213</td>
<td>E (94 %)</td>
<td>hydrogen bond to PLP and interaction with position 169 (R: 50 %, W: 11 %)) (Lyskowski et al., 2014)</td>
</tr>
<tr>
<td>218</td>
<td>N (92 %)</td>
<td></td>
</tr>
<tr>
<td>220</td>
<td>F (76 %), W (7 %), Y (6 %)</td>
<td></td>
</tr>
<tr>
<td>222</td>
<td>V (73 %), I (15 %)</td>
<td></td>
</tr>
<tr>
<td>225</td>
<td>G (77 %), N (7 %), D (5 %)</td>
<td>loop</td>
</tr>
<tr>
<td>229</td>
<td>T (87 %)</td>
<td></td>
</tr>
<tr>
<td>230</td>
<td>P (78 %), R (7 %), H (5 %)</td>
<td></td>
</tr>
<tr>
<td>235</td>
<td>L (94 %)</td>
<td>PLP binding cup (Lyskowski et al., 2014)</td>
</tr>
<tr>
<td>237</td>
<td>G (97 %)</td>
<td>loop</td>
</tr>
<tr>
<td>238</td>
<td>I (81 %), V (10 %)</td>
<td>PLP binding cup (Lyskowski et al., 2014)</td>
</tr>
<tr>
<td>239</td>
<td>T (89 %)</td>
<td>PLP binding cup (Lyskowski et al., 2014)</td>
</tr>
<tr>
<td>240</td>
<td>R (90 %)</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>E (75 %), V (5 %)</td>
<td></td>
</tr>
</tbody>
</table>
A.3 The \(\omega\)-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Conserved amino acids</th>
<th>Location/function</th>
</tr>
</thead>
<tbody>
<tr>
<td>267</td>
<td>A (84 %), F (7 %)</td>
<td></td>
</tr>
<tr>
<td>268</td>
<td>D (72 %), E (9 %)</td>
<td>salt bridge to position 223 (K: 51 %, R: 24 %)</td>
</tr>
<tr>
<td>269</td>
<td>E (94 %)</td>
<td></td>
</tr>
<tr>
<td>271</td>
<td>F (83 %), W (7 %)</td>
<td></td>
</tr>
<tr>
<td>273</td>
<td>T (72 %), S (14 %), C (8 %)</td>
<td></td>
</tr>
<tr>
<td>281</td>
<td>P (84 %), A (8 %)</td>
<td></td>
</tr>
<tr>
<td>286</td>
<td>D (78 %), G (8 %)</td>
<td></td>
</tr>
<tr>
<td>293</td>
<td>G (76 %)</td>
<td>loop</td>
</tr>
<tr>
<td>296</td>
<td>G (89 %)</td>
<td>loop</td>
</tr>
</tbody>
</table>

Table A.9: Substrate specificity-determining positions and substrate-specific sequence motifs in Fold type I (S)-selective \(\omega\)-transaminases. Standard positions refer to position numbers of the \(\omega\)-TA from *Mesorhizobium* sp. LUK (PDB accession 2YKU).

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Position</th>
<th>Function</th>
<th>Source organism</th>
<th>Uniprot ID</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.19</td>
<td>F19W</td>
<td>higher activity towards (\beta)-keto esters</td>
<td><em>V. fluvialis</em></td>
<td>F2XBU9</td>
<td>(Midelfort <em>et al.</em>, 2013)</td>
</tr>
<tr>
<td>26</td>
<td>W57F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53.1</td>
<td>F85A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>V153A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>118.1</td>
<td>K163F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>192</td>
<td>I259V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>346</td>
<td>R415F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53.4</td>
<td>R88K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Position</th>
<th>Function</th>
<th>Source organism</th>
<th>Uniprot ID</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>L20</td>
<td>hydrophobic L-pocket</td>
<td><em>C. crescentus</em> and others</td>
<td>P28269</td>
<td>(Park <em>et al.</em>, 2012)</td>
</tr>
<tr>
<td>0.23</td>
<td>(Y/ W/ L)23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53.2</td>
<td>(Y/ F)88</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>Y152</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.36</td>
<td>R36</td>
<td>decrease of activity towards aromatic $\beta$-amino acid</td>
<td><em>S. thermophilus</em></td>
<td>D1C218</td>
<td>(Mathew <em>et al.</em>, 2016c)</td>
</tr>
<tr>
<td>0.41</td>
<td>R41</td>
<td>coordination of substrate carboxyl-group</td>
<td><em>V. paradoxus</em></td>
<td>H8WR05</td>
<td>(Crismaru <em>et al.</em>, 2013)</td>
</tr>
<tr>
<td>0.41</td>
<td>R41</td>
<td>coordination of substrate carboxyl-group</td>
<td><em>V. paradoxus</em></td>
<td>H8WR05</td>
<td>(Crismaru <em>et al.</em>, 2013)</td>
</tr>
<tr>
<td>0.43</td>
<td>(A/ V/ I)43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>P50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>D65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>G66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>(E/ D/ N/ Q)75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>(Y/ F/ W)76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>G48R</td>
<td>improved stability; activity towards aminotetralin</td>
<td><em>A. citreus</em> (alternative <em>B. megaterium</em>)</td>
<td>A0A1C7D191</td>
<td>(Martin <em>et al.</em>, 2007)</td>
</tr>
<tr>
<td>26</td>
<td>Y60C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Position</th>
<th>Function</th>
<th>Source organism</th>
<th>Uniprot ID</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>118.5</td>
<td>Y164F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>126.5</td>
<td>R186S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>161</td>
<td>A242V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>165</td>
<td>A245T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>172</td>
<td>I252V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>F255I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>188</td>
<td>N268S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>346.9</td>
<td>T409R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>346.24</td>
<td>K424E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>346.36</td>
<td>V436A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>E75</td>
<td>interaction with V. paradoxus</td>
<td>H8WR05</td>
<td>(Crismaru et al., 2013)</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Y85I</td>
<td>shift of activity: Homo sapiens</td>
<td>P04181</td>
<td>(Markova et al., 2005)</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>L56V</td>
<td>increase of activity towards branched chain substrates</td>
<td>V. fluvialis</td>
<td>F2XBU9</td>
<td>(Genz et al., 2016)</td>
</tr>
<tr>
<td>26</td>
<td>W57C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53.1</td>
<td>F85V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>V153A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>L57A</td>
<td>allows for re-face attack; increased activity towards butyrophenone</td>
<td>O. antrhopi</td>
<td>A6WVC6</td>
<td>(Han et al., 2017)</td>
</tr>
<tr>
<td>26</td>
<td>W58A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>W57F</td>
<td>opening of P-pocket</td>
<td>V. fluvialis</td>
<td>F2XBU9</td>
<td>(Midelfort et al., 2013)</td>
</tr>
</tbody>
</table>

©2018 Wiley Periodicals, Inc.
### A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Position</th>
<th>Function</th>
<th>Source organism</th>
<th>Uniprot ID</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>W60C</td>
<td>increase of enantioselectivity and activity towards aromatic substrate</td>
<td><em>C. violaceum</em></td>
<td>Q7NWG4</td>
<td>(Humble <em>et al.</em>, 2012; Deszcz <em>et al.</em>, 2015; Cassimjee <em>et al.</em>, 2012)</td>
</tr>
<tr>
<td>26</td>
<td>W58L</td>
<td>opening of substrate pocket; activity towards aromatic ketones</td>
<td><em>O. anthropi; V. fluvialis</em></td>
<td>A6WVC6</td>
<td>(Midelfort <em>et al.</em>, 2013; Han <em>et al.</em>, 2015)</td>
</tr>
<tr>
<td>26</td>
<td>Y59</td>
<td>determines size of the O-pocket</td>
<td><em>Ruegeria</em> sp. TM1040</td>
<td>Q1GD43</td>
<td>(Pavlidis <em>et al.</em>, 2016; Steffen-Munsberg <em>et al.</em>, 2013)</td>
</tr>
<tr>
<td>26</td>
<td>W60</td>
<td>determines size of the S-pocket</td>
<td><em>C. crescentus</em> P28269</td>
<td>(reviewed by Park <em>et al.</em>, 2012)</td>
<td></td>
</tr>
<tr>
<td>161</td>
<td>S231</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>192</td>
<td>I262</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>G98M</td>
<td>increase of stability</td>
<td><em>V. paradoxus</em></td>
<td>4AOA</td>
<td>(Buß <em>et al.</em>, 2017)</td>
</tr>
<tr>
<td>53.1</td>
<td>Y87</td>
<td>interacts with aromatic substrate in P-pocket</td>
<td><em>Ruegeria</em> sp. TM1040</td>
<td>Q1GD43</td>
<td>(Pavlidis <em>et al.</em>, 2016; Steffen-Munsberg <em>et al.</em>, 2013)</td>
</tr>
<tr>
<td>53.1</td>
<td>F85L</td>
<td>activity towards PEA and longer side chains</td>
<td><em>V. fluvialis</em></td>
<td>F2XBU9</td>
<td>(Midelfort <em>et al.</em>, 2013; Nobili <em>et al.</em>, 2015a)</td>
</tr>
<tr>
<td>111</td>
<td>V153A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Position</th>
<th>Function</th>
<th>Source organism</th>
<th>Uniprot ID</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>53.1</td>
<td>F85A</td>
<td>increase of binding pocket</td>
<td><em>V. fluvialis</em></td>
<td>F2XBU9</td>
<td>(Midelfort et al., 2013)</td>
</tr>
<tr>
<td>53.1</td>
<td>F88A</td>
<td>inversion of enantiopreference from <em>(S)</em> to <em>(R)</em></td>
<td><em>C. violaceum</em></td>
<td>Q7NWG4</td>
<td>(Humble et al., 2012)</td>
</tr>
<tr>
<td>53.2</td>
<td>F92V</td>
<td>inhibits activity towards aromatic PEA</td>
<td><em>Ruegeria</em> sp.</td>
<td>Q5LMU1</td>
<td>(Steffen-Munsberg et al., 2013)</td>
</tr>
<tr>
<td>108</td>
<td>Y153M/S/N</td>
<td>switch from a α-TA to a ω-TA</td>
<td><em>C. violaceum</em></td>
<td>Q7NWG4</td>
<td>(Deszcz et al., 2015)</td>
</tr>
<tr>
<td>108</td>
<td>Y150F</td>
<td>higher activity towards amino alcohols</td>
<td><em>V. fluvialis</em></td>
<td>F2XBU9</td>
<td>(Nobili et al., 2015b)</td>
</tr>
<tr>
<td>111</td>
<td>V153A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>108</td>
<td>Y152</td>
<td>determines size of the small pocket and allows only methyl residue of PEA</td>
<td><em>P. putida</em></td>
<td>P28269</td>
<td>(Park et al., 2012)</td>
</tr>
<tr>
<td>192</td>
<td>I262</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>V153A</td>
<td>increases size of P-pocket of ω-TA; activity towards aliphatic α-keto acids</td>
<td><em>P. denitrificans</em></td>
<td>A1B956</td>
<td>(Park et al., 2014)</td>
</tr>
<tr>
<td>118.1</td>
<td>N161I</td>
<td>improved stability</td>
<td><em>P. mandelii</em></td>
<td>A0A059KSX8</td>
<td>(Börner et al., 2017)</td>
</tr>
<tr>
<td>118.5</td>
<td>Y164L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>V228G</td>
<td>increase of activity towards aromatic β-amino acid</td>
<td><em>C. crescentus</em></td>
<td>Q7WWK8</td>
<td>(Hwang et al., 2008)</td>
</tr>
</tbody>
</table>
### A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Position</th>
<th>Function</th>
<th>Source organism</th>
<th>Uniprot ID</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>220</td>
<td>N286A</td>
<td>tolerance for alcohol ester substrate</td>
<td><em>V. fluvialis</em></td>
<td>F2XBU9</td>
<td>(Midelfort <em>et al.</em>, 2013)</td>
</tr>
<tr>
<td>192</td>
<td>I259V</td>
<td>Inversion of enantiopreference from (S) to (R)</td>
<td><em>A. citreus</em></td>
<td>A0A1C7D191</td>
<td>(Svedendahl <em>et al.</em>, 2010)</td>
</tr>
<tr>
<td>248</td>
<td>V328A</td>
<td>Increases enantiopreference for (S)</td>
<td><em>A. citreus</em></td>
<td>A0A1C7D191</td>
<td>(Svedendahl <em>et al.</em>, 2010)</td>
</tr>
<tr>
<td>251</td>
<td>Y331C</td>
<td>less polarity inside P-pocket</td>
<td><em>V. fluvialis</em></td>
<td>F2XBU9</td>
<td>(Midelfort <em>et al.</em>, 2013)</td>
</tr>
<tr>
<td>346</td>
<td>R415</td>
<td>flipping arginine (dual substrate recognition)</td>
<td><em>V. fluvialis</em></td>
<td>F2XBU9</td>
<td>(Midelfort <em>et al.</em>, 2013)</td>
</tr>
<tr>
<td>346</td>
<td>R415F</td>
<td>loss of activity P-pocket</td>
<td><em>C. crescentus</em></td>
<td>P28269</td>
<td>(Park <em>et al.</em>, 2014)</td>
</tr>
<tr>
<td>346.1</td>
<td>R414</td>
<td>flipping arginine ($\alpha$-carboxyl binding site)</td>
<td><em>Pseudomonas</em> sp. strain AAC</td>
<td>A0A081YAY5</td>
<td>(Wilding <em>et al.</em>, 2016)</td>
</tr>
<tr>
<td>346.4</td>
<td>P423</td>
<td>entrance of substrate pocket</td>
<td><em>Ruegeria</em> sp. TM1040</td>
<td>Q1GD43</td>
<td>(Pavlidis <em>et al.</em>, 2016)</td>
</tr>
<tr>
<td>346.20</td>
<td>R416</td>
<td>flipping arginine (dual substrate recognition)</td>
<td><em>C. violaceum</em></td>
<td>Q7NWG4</td>
<td>(Manta <em>et al.</em>, 2017)</td>
</tr>
</tbody>
</table>
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Table A.10: Substrate specificity-determining positions and substrate-specific sequence motifs in Fold type IV ($R$)-selective $\omega$-transaminases. Standard positions refer to position numbers of the $\omega$-TA from *Aspergillus terreus* (PDB accession 4CE5).

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Position</th>
<th>Function</th>
<th>Source organism</th>
<th>Uniprot ID</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>H53</td>
<td>specificity towards $(R)$-amines</td>
<td><em>Nectria haemato-cocca</em></td>
<td>C7YVL8</td>
<td>(Sayer <em>et al.</em>, 2014)</td>
</tr>
<tr>
<td>60</td>
<td>Y58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>V60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>H62A</td>
<td>increase of activity towards aromatic ketone substrate</td>
<td><em>Arthrobacter</em> 117</td>
<td>F7J696-1</td>
<td>(Savile <em>et al.</em>, 2010)</td>
</tr>
<tr>
<td>62</td>
<td>V69G</td>
<td>increases size of small pocket</td>
<td><em>Arthrobacter</em> 117</td>
<td>F7J696-1</td>
<td>(Savile <em>et al.</em>, 2010)</td>
</tr>
<tr>
<td>115</td>
<td>F122I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>276</td>
<td>A284G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>E125</td>
<td>entrance tunnel limiting the substrate size</td>
<td><em>C.pusillum</em></td>
<td>A0A1S4NYF0</td>
<td>(Pavkov-Keller <em>et al.</em>, 2016)</td>
</tr>
<tr>
<td>147</td>
<td>E140</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>G136Y</td>
<td>increase of hydrophobic interaction with substrate</td>
<td><em>Arthrobacter</em> 117</td>
<td>F7J696-1</td>
<td>(Savile <em>et al.</em>, 2010)</td>
</tr>
<tr>
<td>127</td>
<td>E137I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>191</td>
<td>V199I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>201</td>
<td>A209L</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>R138</td>
<td>interacting with keto group of the substrate</td>
<td><em>Arthrobacter</em> 117</td>
<td>F7J696-1</td>
<td>(Savile <em>et al.</em>, 2010)</td>
</tr>
</tbody>
</table>
### A.3 The ω-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

<table>
<thead>
<tr>
<th>Standard position</th>
<th>Position</th>
<th>Function</th>
<th>Source organism</th>
<th>Uniprot ID</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>T130M</td>
<td>increase of thermostability</td>
<td>Aspergillus terreus</td>
<td>Q0C8G1</td>
<td>(Huang et al., 2017)</td>
</tr>
<tr>
<td>133</td>
<td>E133F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>215</td>
<td>S223P</td>
<td>increases size of large pocket</td>
<td>Arthrobacter</td>
<td>F7J696-1</td>
<td>(Savile et al., 2010)</td>
</tr>
<tr>
<td>274</td>
<td>T273</td>
<td>enantiopreference by limiting space in small pocket</td>
<td>Nectria haematococca</td>
<td>C7YVL8</td>
<td>(Sayer et al., 2014)</td>
</tr>
<tr>
<td>275</td>
<td>T274</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>276</td>
<td>A275</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A.3.6 Supporting Information


Figure A.25: Distribution of sequence lengths of Fold type IV (oTA IV) and Fold type I (oTA I), ω-transaminases from the oTAED, represented as histograms and boxplots with maximum whisker length 1.5 times the interquartile range and outliers marked as crosses.
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Figure A.26: $\omega$-transaminases of Fold type I (A) and Fold type IV (B) from psychrophilic (green), thermophilic (red), and hyperthermophilic (black) sources retrieved from the BacDive database (release 27.02.2017) (Söhngen et al., 2014) and the representative node (orange) containing the motif from Stekhanova et al. (2017) for thermostable Fold type IV sequences (V/I)xLDxR and PFG(K/H)YL. The nodes correspond to representative sequences of clusters formed by 30% identity in USEARCH (Edgar, 2010). A cutoff of 50% pairwise sequence similarity was used to construct the edges. The network is shown as force-directed layout, thus pairs with higher similarity are positioned in closer proximity.
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Table A.11: Query sequences for the initial BLAST search to set up the $\omega$-Transaminase Engineering Database (oTAED).

<table>
<thead>
<tr>
<th>UniProt identifier</th>
<th>Annotation</th>
<th>Source organism</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>H8WR05</td>
<td>Beta-Phenylalanine Aminotransferase</td>
<td>Variovorax paradoxus</td>
<td>(Crismaru et al., 2013)</td>
</tr>
<tr>
<td>Q0CSG1</td>
<td>AT-OmegaTA Multispecies: Aminotransferase IV</td>
<td>Aspergillus terreus</td>
<td>(Lyskowski et al., 2014)</td>
</tr>
<tr>
<td>A1TD51</td>
<td>Multispecies: Aminotransferase IV</td>
<td>Mycobacterium vanbaalenii</td>
<td>(Mathew et al., 2016b)</td>
</tr>
<tr>
<td>A1DD33</td>
<td>Aminotransferase, Aspergillus fischer</td>
<td></td>
<td>(Mathew et al., 2016b)</td>
</tr>
<tr>
<td>F2XBU9</td>
<td>Pyruvate transaminase</td>
<td>Vibrio fluvialis</td>
<td>(Genz et al., 2016)</td>
</tr>
<tr>
<td>A0A1B4YG16</td>
<td>Beta-Alanine Pyruvate Transaminase</td>
<td>Mesorhizobium loti</td>
<td>(Shin et al., 2003)</td>
</tr>
<tr>
<td>P28269</td>
<td>Omega Amino Acid: Pyruvate Aminotransferase</td>
<td>Pseudomonas putida</td>
<td>(Yonaha et al., 1983)</td>
</tr>
<tr>
<td>D5VI64</td>
<td>Aspartate Aminotransferase Family Protein</td>
<td>Caulobacter segnis</td>
<td>(Sayer et al., 2013)</td>
</tr>
<tr>
<td>Q9I700</td>
<td>Aspartate Aminotransferase Family Protein</td>
<td>Pseudomonas aeruginosa</td>
<td>(Höhne et al., 2010)</td>
</tr>
</tbody>
</table>
A.3 The $\omega$-transaminase engineering database (oTAED): a navigation tool in protein sequence and structure space

Table A.12: Reference structures for the standard numbering schemes of $\omega$-transaminase superfamilies Fold Type I and Fold Type IV.

<table>
<thead>
<tr>
<th>Fold Type I PDB entry</th>
<th>Fold Type I Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1DTY</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>1SZS</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>2OAT</td>
<td><em>Homo sapiens</em></td>
</tr>
<tr>
<td>2YKU</td>
<td><em>Mesorhizobium</em> sp. LUK</td>
</tr>
<tr>
<td>2ZSL</td>
<td><em>Aeropyrum pernix</em></td>
</tr>
<tr>
<td>3A8U</td>
<td><em>Pseudomonas putida</em></td>
</tr>
<tr>
<td>3DOD</td>
<td><em>Bacillus subtilis</em></td>
</tr>
<tr>
<td>3FCR</td>
<td><em>Ruegeria</em> sp. TM1040</td>
</tr>
<tr>
<td>3I5T</td>
<td><em>Rhodobacter sphaeroides</em></td>
</tr>
<tr>
<td>4AOA</td>
<td><em>Variovorax paradoxus</em></td>
</tr>
<tr>
<td>4E3R</td>
<td><em>Homo sapiens</em></td>
</tr>
<tr>
<td>4JEW</td>
<td><em>Salmonella enterica</em></td>
</tr>
<tr>
<td>4UOX</td>
<td><em>Escherichia coli</em></td>
</tr>
<tr>
<td>4YSN</td>
<td><em>Lactobacillus buchneri</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fold Type IV PDB entry</th>
<th>Fold Type IV Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>3WWH</td>
<td><em>Arthrobacter</em> sp. KNK168</td>
</tr>
<tr>
<td>4CE5</td>
<td><em>Aspergillus terreus</em></td>
</tr>
<tr>
<td>4CHI</td>
<td><em>Aspergillus fumigatus</em></td>
</tr>
<tr>
<td>4CMD</td>
<td><em>Nectria haematococca</em></td>
</tr>
<tr>
<td>5E25</td>
<td><em>Geoglobus acetivorans</em></td>
</tr>
<tr>
<td>5K3W</td>
<td><em>Curtobacterium pusillum</em></td>
</tr>
</tbody>
</table>

Table A.13: Relevant fingerprints of oTAED Fold Type IV standard positions as reported previously (Höhne et al., 2010) for 4-amino-4-deoxychorismate lyase (ADCL), D-amino acid aminotransferases (DATA), L-branched chain amino acid aminotransferases (L-BCAT) and (R)-amine transaminases (ATA). In addition, the catalytic lysine at Fold Type IV standard position 180 is conserved in the four subfamilies.

<table>
<thead>
<tr>
<th>Standard position</th>
<th>55</th>
<th>60</th>
<th>62</th>
<th>64</th>
<th>115-</th>
<th>128-130</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADCL</td>
<td>F/Y</td>
<td>F</td>
<td>T/S</td>
<td>-</td>
<td>(V/I/L)x(K/R)</td>
<td>RGY</td>
</tr>
<tr>
<td>DATA</td>
<td>F</td>
<td>Y/E/D</td>
<td>V</td>
<td>K/R</td>
<td>x(V/I/L)Y(V/I/L)Q</td>
<td>RxH</td>
</tr>
<tr>
<td>L-BCAT</td>
<td>Y</td>
<td>F/E/D</td>
<td>G</td>
<td>R/K</td>
<td>Y(V/I/L)R</td>
<td>(V/I/L)G(V/I/L)</td>
</tr>
<tr>
<td>ATA</td>
<td>H/R</td>
<td>Y</td>
<td>V/T</td>
<td>S/T/A/H/P</td>
<td>(F/Y)V(E/A/S/N/Q)</td>
<td>-</td>
</tr>
</tbody>
</table>
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments


Own contributions

I refined the relational data model of BioCatNet (see chapter 2.2.1) and curated the data provided by experimental collaborators. Based on the previous work by Reusch (2014) and Vogel (2015) I conceptualized BioCatNet with a focus on time-course data for kinetic modeling in upcoming projects. I contributed to the writing of the original manuscript.

A.4.1 Abstract

The development of novel enzymes for biocatalytic processes requires knowledge on substrate profile and selectivity which can be derived from separate databases and publications. Often, these data sources lack the time courses of substrate or product, and an unambiguous link between experiment and enzyme sequence. The lack of integrated, original data hampers the comprehensive analysis of enzyme kinetics and the evaluation of sequence-function relationships. To accelerate enzyme engineering, BioCatNet integrates protein sequence, structure, and original experimental data for a given enzyme family. BioCatNet explicitly assigns the enzyme sequence to an experiment, which consists of information on reaction conditions and time course data. BioCatNet facilitates the consistent documentation of reaction conditions, the archiving of measured time course data, and the efficient exchange of original experimental data among collaborators. Data integration is demonstrated for three case studies using the Thiamine diphosphate-dependent Enzymes Engineering Database\textsuperscript{15}.

A.4.2 Introduction

Enzymes are widely used as biocatalysts with desirable activity and selectivity for an increasing range of interesting synthetic applications (Adrio and Demain, 2014). Since the availability of known, natural enzymes that perform in a given process is still limited, there is an urgent need for efficient methods to identify new enzyme candidates and to improve enzymes by enzyme engineering, especially for non-natural reactions. The development of new enzymes aims at desired properties such as optimized catalytic activity, solvent stability, chemo-, regio- or stereoselectivity (Bornscheuer et al., 2012). The efficient development of new enzymes by computa-

\textsuperscript{15} https://teed.biocatnet.de, accessed on April 30, 2018.
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

The development of computational methods such as data mining and molecular modeling is based on a profound knowledge of sequence-structure-function relationships (Bornscheuer, 2016). Consequently, data on protein sequences, structures, and the effects of mutations on biochemical properties have become precious raw materials for successful protein design (Pleiss, 2014). Comprehensive online repositories such as the National Center for Biotechnology Information (NCBI) Protein Database (Benson et al., 2011), UniProt (Apweiler et al., 2014), DNA Data Bank of Japan (DDBJ) (Kaminuma et al., 2011), and Protein Data Bank (PDB) (Berman et al., 2000) enable systematic analyses of sequences and structures of biocatalysts, while the experimental results from mutational studies are reported in scientific papers and are accessible by repositories such as NCBI PubMed16, Web of Science17, Google Scholar18 or SciFinder19.

Systematic analysis of available protein sequence and structure data allows for identification of functionally relevant amino acids and the prediction of promising mutation sites by determining conserved (Lehmann et al., 2002; Anbar et al., 2012), family-specific (Mazin et al., 2010; Suplatov et al., 2012, 2014), or correlated positions (Kuipers et al., 2009; Kowarsch et al., 2010). To enable the analysis of a rapidly increasing number of protein sequences and structures, enzyme family databases have been established for cytochrome P450 monooxygenases (Fischer et al., 2007), β-lactamas (Widmann et al., 2012), triterpene-cyclases (Racolta et al., 2012), imine reductases (Scheller et al., 2014), lipases and other α/β-hydrolases (Fischer and Pleiss, 2003; Hotelier et al., 2004), carbohydrate-active enzymes (Lombard et al., 2013), and thiamine diphosphate-dependent enzymes (Widmann et al., 2010). The commercially available 3DM databases follow a similar approach by using multiple structure alignments of reference enzymes to describe an enzyme family (Kuipers et al., 2010b). Among others, the application of 3DM databases includes the automated assignment of mutations from literature sources (Kuipers et al., 2010a). Apart from enzyme family databases, the UniProt database contains information on natural variants of proteins (Apweiler et al., 2014).

Although family databases make the rapidly increasing number of protein sequences and structures accessible to systematic analysis, they lack biocatalytic experimental data. Providing consistently linked data on sequence, structure, and experimentally determined biochemical properties of enzyme variants will result in invaluable information resources for directly deriving sequence-function relationships.

Experimental characterization of enzymes and designed variants results in a large quantity of highly heterogeneous data, which is usually reported as plain text, tables or figures in publications or laboratory notebooks. It would be desirable if the information on the enzyme, the reaction conditions, and the biochemical properties were integrated and accessible to an auto-

A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

mated analysis. In most studies systematic experiments have been performed by varying the sequence of the biocatalyst, the substrate scope, or the reaction conditions, and by measuring the respective catalytic activity, chemo-, regio- or stereoselectivity, but only a subset of most relevant data was selected for publication, which results in a substantial loss of information. Since scientific publications are the only available source of experimental data on sequence-function relationships, a scientist has to invest considerable time in finding, collecting, and combining this fragmented and sparse data. To support this data collection process, databases such as the Braunschweig Enzyme Database (BRENDA) (Schomburg et al., 2012), ExplorEnz (McDonald et al., 2007), Kyoto Encyclopedia of Genes and Genomes (KEGG) REACTION (Kanehisa et al., 2006), and SABIO-RK (Wittig et al., 2012) incorporate biochemical information obtained by literature mining and provide comprehensive repositories of enzyme function. However, these databases lack unambiguous information on the complete protein sequence of an enzyme as well as information on the reaction conditions and the kinetic models which were used to derive the reported kinetic parameters.

To fill this gap, community-wide standards on reporting enzymatic data (Tipton et al., 2014; Gardossi et al., 2010) and comprehensive databases on biocatalytic experiments have been proposed (Apweiler et al., 2010). Minimal requirements for reporting biocatalytic results were suggested by the Standards for Reporting Enzyme Data (STRENDA) Consortium20 (Tipton et al., 2014). These recommendations include an unambiguous identification of the applied enzyme, the reaction conditions and the experimental outcome.

However, all currently available databases report only kinetic parameters such as $K_m$ and $k_{cat}$ rather than the underlying time course data monitoring the conversion of substrate(s) to product(s). While the kinetic parameters are a consequence of the assumed kinetic model, reporting the underlying time course data would allow for the comparison of different kinetic models to characterize biochemical function (Zavrel et al., 2008; Schnell, 2014). For a model-independent documentation of biochemical function, it would be crucial to report original data on substrate conversion or product formation rather than kinetic parameters only. The combination of time course data, a documentation of reaction conditions, and an unambiguous link to the respective protein sequence would considerably improve the reproducibility of published experiments and allow for systematic computational analyses of sequence-function relationships. Therefore, we established the BioCatNet system which integrates sequence and structure of enzymes with biocatalytic data such as reaction conditions and time-dependent biochemical data (Figure A.27). Besides substrate or product concentration, BioCatNet can store conversion, yield or enantiomeric excess, since these types of data are usually generated upon screening potential enzyme candidates for activity or selectivity towards various substrates.

Application of BioCatNet is demonstrated on reactions catalyzed by thiamine diphosphate

A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

(ThDP)-dependent enzymes. Therefore, sequence and structure data from the TEED, the ThDP-dependent Enzymes Engineering Database (Widmann et al., 2010), have been migrated to the BioCatNet system, updated, and enriched by experimentally and computationally derived information. The BioCatNet system can be used as a comprehensive platform to store and exchange biochemical data on any protein family and to make this data accessible to selected users for later systematic analyses.

A.4.3 Results and Discussion

The BioCatNet Concept The BioCatNet database system was established as a repository to analyze the sequence-function relationships of a complete enzyme family. Original data is parsed from data repositories or experiments, such as sequences, structures or biocatalytic measurements (Figure A.28). Data analysis is based on mathematical models and therefore produces ”model-derived information”. Examples are sequence alignments, which are derived from original sequence data by applying a distance metric and alignment parameters, and kinetic parameters, which are derived from original time course data by applying a kinetic model.

Original sequence data is parsed from the NCBI protein database by BLAST (Basic Local Alignment Search Tool) searches starting from a set of seed sequences (Fischer et al., 2006). Model-derived information for sequence data comprises hierarchical sorting of individual sequences to proteins and homologous families by the USEARCH clustering algorithm (Edgar, 2010), sequence alignments generated by Clustal Omega (Sievers et al., 2011), generation of family-specific sequence profiles by profile hidden Markov models (Eddy, 1998), assignment of standard position numbers (Vogel et al., 2012), links to literature entries (Gricman et al., 2015), and the annotation of structurally or functionally relevant positions. Additionally, structure data can be used for homology modeling.

Experimental data is provided by experimenters rather than parsed automatically from literature sources. Each experimental data set refers unambiguously to the protein sequence of the respective enzyme. BioCatNet has been designed as a tool for experimenters to store, organize, retrieve, and analyze their own experimental data or data from collaborative projects. Experimental data include the reaction conditions as suggested by the STRENSA Consortium (Tipton et al., 2014), as well as time-dependent data of substrate and product concentrations or of derived properties such as enantiomeric excess (ee values), conversions, or yields (Table A.14).

As an initial example of an enzyme family database, the TEED (Widmann et al., 2010) was migrated to BioCatNet21 (Vogel, 2015).

21 See footnote 15 on page 174.
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

**Reporting of Experimental Data** The minimal requirements for submission of experimental measurement data to BioCatNet constitute a compromise between completeness and usability Table A.14. Therefore, BioCatNet distinguishes between mandatory and optional attributes for data submission. Since the description of the reaction under investigation was considered as indispensable to map sequence-function relationships, the involved substrates and products as well as the reaction stoichiometry are mandatory inputs. To describe the reaction conditions, the specification of the reaction buffer is mandatory, while additional information such as pH and temperature are optional but recommended inputs. All additional information on preparation of enzymes, analytical methods and other laboratory routines are considered not to be mandatory, but are usually documented in laboratory journals or by a laboratory information management system (LIMS).

Before experimental measurement data can be uploaded, BioCatNet requires the distinct protein sequence of the respective enzyme, which guarantees an unambiguous distinction between the biocatalytic properties of enzyme variants or homologues. We recommend the mentioning of source organism and expression host for a given enzyme. The usage of the SMILES code (simplified molecular-input line-entry system) (Weininger, 1988) is recommended as unambiguous identifier of a substrate or product, which facilitates the search for experimental data by compounds. In order to ease usage of the SMILES code, the BioCatNet documentation\(^\text{22}\) links to the SMILES generator NCBI PubChem Sketcher (Ihlenfeldt et al., 2009). Substrates, enzymes, and additive compounds can be quantified either by a combination of amount of substance and volume or by concentration.

To facilitate submission of biocatalytic data to BioCatNet, a Microsoft Excel template is provided within the documentation page\(^\text{23}\). The following examples describe case studies for experiments combining a substrate with different enzyme variants, measurement of time courses for substrate and product concentrations, and conversion of substrates under varying conditions.

**Example 1: Experiments with Different Substrates and Enzyme Variants** As a first example, biocatalytic experiments using C-terminally His-tagged pyruvate decarboxylase from *Acetobacter pasteurianus* (*ApPDC*) and the variant E469G (Rother neé Gocke et al., 2011; Rother et al., 2014; Sehl et al., 2017) were uploaded to the TEED. The sequence of the *ApPDC* variant was uploaded, and annotations of mutated residues, domain boundaries of the catalytically relevant pyrophosphate- and pyrimidine-interacting domains, and the standard position numbers for ThDP-dependent decarboxylases (Vogel et al., 2012) were generated, thus facilitating the communication of mutation sites and the analysis of conserved positions.


\(^{23}\) See footnote 8 on page 51.
In the biocatalytic experiments, the ApPDC-catalyzed coupling reactions of benzaldehyde with pyruvate was monitored by measuring conversion of benzaldehyde and enantiomeric excess (ee) of phenylacetylcarbinol after a reaction time of 16 hours. Each experiment describes the measurement using one enzyme. For each experiment, the enzymatic reaction, the reaction conditions including a description of the reaction buffer and the enzyme preparation, the substrate, and the enzyme sequence are specified. The two experiments were grouped into a single ”experiment set”. Reaction conditions and experimental results are visualized as tables in the web interface (Figure A.30). A summary of the experimental setup can be displayed as full-text for a comprehensive documentation of the biocatalytic experiments. Excel files for the two experiments were created by filling in the Excel template (Supporting Information and) and can be downloaded from the documentation page.

Example 2: Experiments with Measured Time Courses of Substrate and Product Concentration  Time-dependent substrate and product concentrations upon self-ligation of benzaldehyde to benzoin catalyzed by benzaldehyde lyase from Pseudomonas fluorescens (PfBAL) were measured for varying initial concentrations of substrate and product (Ohs et al., 2018). The C-terminally His-tagged sequence, reaction conditions, and the time course data of one experiment were uploaded to the TEED. The time course data are visualized on the web interface (Figure A.31). An Excel file of the experiment is available as supporting information and can be downloaded from the documentation page.

Example 3: Experiments under Varying Conditions  The conversion of benzaldehyde and acetaldehyde to (S)-2-hydroxy-1-phenylpropan-1-one was catalyzed by C-terminally His-tagged benzoylformate decarboxylase under varying pH to find the optimum pH for the recently described ”µMORE” microfluidics reactor (Jussen et al., 2016). The time courses of benzaldehyde and (S)-2-hydroxy-1-phenylpropan-1-one were monitored in an ”experiment set” for the pH values of 6.5 and 8.0. Two independent experiments were conducted for each pH value. While a publication usually contains only mean values and standard deviations, BioCatNet keeps all original time courses in a readable format for future analyses. The storage of original time course data allows for a model-based analysis of multiple data sets to study the effect of reaction conditions (pH, temperature) on the reaction kinetics, which might later serve as a basis for reaction optimization process design. Excel files of the experiment set are available as

---

24 wiki.biocatnet.de/wiki/examples/example1a.xlsx, accessed on April 30, 2018.
26 See footnote 22 on page 178.
27 wiki.biocatnet.de/wiki/examples/example2.xlsx, accessed on April 30, 2018.
28 See footnote 22 on page 178.
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

supporting information\textsuperscript{29}, \textsuperscript{30}, \textsuperscript{31}, \textsuperscript{32}, and can be downloaded from the documentation page\textsuperscript{33}.

The Benefit of Unambiguously Linking Sequence and Biochemical Function

The protein sequence uniquely identifies an enzyme. Nevertheless, in many publications the respective enzyme is specified ambiguously by naming its enzymatic function and its source organism, e.g. ”pyruvate decarboxylase from \textit{Acetobacter pasteurianus}”. A search for this term in the NCBI protein database matches 30 protein entries, which differ not only in length but also in positions distributed over the entire protein sequence (Supporting Information Figure A.33).

While highly similar sequences commonly fold into similar three-dimensional structures (Browne \textit{et al}., 1969) and usually resemble each other in their catalytic function (Chattopadhyay \textit{et al}., 2009), single mutations might affect the activity and specificity, even if the mutated positions are outside the active site (Drawz \textit{et al}., 2009). Specifically, the effect of protein fusion tags for affinity chromatography on activity, stability, and selectivity of an enzyme is often neglected in scientific publications (Majorek \textit{et al}., 2014). Thus, protein sequence, structure, and experimental results have to be assigned consistently in order to enable a comprehensive investigation of the sequence-function relationships of a biocatalyst.

Representation of Biochemical Data

It has been widely recognized that detailed results from biocatalytic experiments are an invaluable source of data for many fields of research, including protein design (Adrio and Demain, 2014; Bornscheuer \textit{et al}., 2012; Bornscheuer, 2016; Pleiss, 2014), synthetic biochemical pathway construction (Bhan \textit{et al}., 2013), reaction optimization and process design. Therefore, guidelines for reporting enzymatic (Tipton \textit{et al}., 2014) and biocatalytic data (Gardossi \textit{et al}., 2010) have been proposed. Guidelines have also been proposed in other fields of research such as systems biology (Schilling \textit{et al}., 2008) and crystalization (Raccuglia \textit{et al}., 2016). Storing data in the BioCatNet system or in the Excel files facilitates the collaboration of different research groups ranging from biocatalysis to modeling.

In BioCatNet, biocatalytic data is represented either as time courses of substrate or product concentrations or as time-dependent ratios such as conversion, yield or enantiomeric excess, depending on the desired type of experiment. The availability of original data circumvents the limitation of data retrieved manually from scientific papers and facilitates documentation of results. For many published kinetic parameters, the underlying original data or the detailed reaction conditions are not available, hindering in-depth analysis of kinetic properties by dif-
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

different kinetic models. The widely used Michaelis-Menten equation is only valid under specific conditions (Apweiler et al., 2010) and excludes enzyme inhibition or inactivation by substrate or product. The lack of time course data results from difficulties in data generation, especially in former studies when methods like high performance liquid chromatography (HPLC) or fluorimeters where yet unavailable. We suggest to make time-resolved data accessible, whenever they have been determined. Besides time course data, ratios like conversion, yield and enantiomeric excess allow for computational comparison of potential enzyme candidates, as previously shown for data on activity and selectivity for cyclododecanone monooxygenase (Fink et al., 2012).

A.4.4 Conclusion

The availability of original time course data enables researchers to determine kinetic parameters and to explore different enzymatic mechanisms (Zavrel et al., 2008). Kinetic modeling provides a link between individual biochemical experiments, and bioprocess design facilitates the search for productive reaction conditions to improve the performance of a biocatalyst and the overall process in a bioreactor (Vasić-Rački et al., 2003). In this context, the BioCatNet system supports both the design of enzymes and the optimization of reaction parameters by linking sequence, structure and biochemical data of an enzyme family within collaborative projects. While the current version of BioCatNet describes reactions catalyzed by a single enzyme, the data model will be extended in the future to describe more complex reaction schemes such as pre-incubation or cascade reactions.

A.4.5 Computational Methods

The BioCatNet Data Model The BioCatNet data model integrates original data such as protein sequence, protein structure, reactions conditions, and time course data (concentrations or derived properties as a function of time). BioCatNet uses the relational database management system Firebird 34. The data model of the DWARF system (Fischer et al., 2006) was extended by tables for taxonomic relations, experimental data, and tables for user management.

The DWARF model organizes sequences in four hierarchical levels of sequences, proteins, homologous families, and superfamilies. In BioCatNet, groups of homologous families and groups of superfamilies were added to the hierarchy to allow for additional sorting by biological knowledge like similarities in domain arrangement and annotation. For each sequence, individual positions can be annotated in BioCatNet by standard positions (Vogel et al., 2012) or function. The data model of the DWARF system was amended in BioCatNet for taxonomic information about the biological source, including taxonomic synonyms, represented by links to NCBI taxonomy.

34 See footnote 9 on page 51
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

Additional tables were added for data on the reaction under investigation, reaction conditions, compounds, and time course data. The link between the respective sequence entry of the experimentally applied enzyme and the experimental time course data was established by a dedicated table (Figure A.32). In addition, the data model was enabled to cope with compounds that are neither substrates nor products, such as inhibitors or detergents. To allow for cascade reactions, a single experiment can include multiple enzymes. For fed-batch experiments, the time points of enzyme or substrate addition can be specified.

The BioCatNet Web Interface The web-accessible user interface for BioCatNet was developed in PHP 5\(^{35}\) (backend) and HTML5/JavaScript (frontend) and is hosted on an Apache webserver\(^{36}\). While sequence and structure data usually is accessible to all users, access to individual biocatalytic data sets is limited to registered users or user groups. Users can navigate between the hierarchical levels of sequence information and taxonomic lineage. Sequences are available as FASTA files and as alignments of homologous families or superfamilies. Experimentally determined structural data and homology models are presented using PV protein viewer\(^{37}\). Biocatalytic data is grouped in ”experiments” and ”experiment sets”. Currently, the upload of the sequences and of experimental data will be done by the BioCatNet administrator in the framework of collaborative projects. In the future, we plan to allow direct upload via a web form to registered users. Furthermore, we would also encourage the use of our Excel template for documentation and communication of biocatalytic data outside BioCatNet. A documentation of the interface and the data model is available online\(^{38}\).

Acknowledgements

The authors thank Anna Baierl, Maryam Beigi, Saskia Bock, Alexander Fries, Sabrina Loschon-sky, Michael Müller, Rüdiger Ohs, Simon Waltzer, and Robert Westphal for valuable discussion on the data model and the necessary standardization in biocatalysis, as well as helpful suggestions concerning the user interface. Support of Michael Widmann in the early phase of the development of the BioCatNet system and suggestions by Catharina Zeil are gratefully acknowledged. This work was funded by the Deutsche Forschungsgemeinschaft within the framework of FOR 1296.

\(^{38}\) See footnote 22 on page 178.
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

Figure A.27: Workflow between bioinformatics and experiment using the BioCatNet database system. Arrows pointing from left to right represent data analysis or modeling, arrows pointing from right to left represent submission of biocatalytic data via file upload. Dashed arrows mark initial procedures and modeling steps beyond the scope of the database.
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

Figure A.28: Concept of the established BioCatNet database system. Original data are stored in the individual enzyme family database on the levels of protein sequence, structure and experimental data. Analysis of data leads to new knowledge which helps to amend the database (model-derived information). Sequences and structures are parsed from other repositories whereas experimenters contribute experimental data and sequences of their applied enzymes. Expansions of the workflow are currently under development to include metagenomic sequences and establish a framework for estimation of kinetic parameters (*).

Figure A.29: Screenshot detail from the BioCatNet web interface depicting a C-terminally His-tagged variant of pyruvate decarboxylase from *Acetobacter pasteurianus*. The Mutated residue (G469), domain boundaries (PYR: pyrimidine interaction, PP: pyrophosphate interaction) and the His-tag are annotated. "DC" indicates the respective standard position number (Vogel *et al.*, 2012).
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

Reaction Conditions

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mM EPPS-buffer with 0.1 mM THF and 2.5 Mgl+</td>
<td></td>
</tr>
<tr>
<td>Experiment was conducted using purified enzyme.</td>
<td></td>
</tr>
<tr>
<td>Shaking Frequency (rpm)</td>
<td>600</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>20</td>
</tr>
</tbody>
</table>

Reaction Buffer

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPPS-buf</td>
<td>potassium phosphate buffer</td>
</tr>
</tbody>
</table>

The reaction from benzaldehyde to benzoin was investigated.
The applied biocatalyst, annotated as ‘Phytoreduced Dehydrogenase from Aspergillus niger 64890/I445’ (EHM1), was expressed in Escherichia coli BL21(DE3) and used as purified enzyme (90% with 7% insoluble, 1% soluble).
The reaction was performed using 50 mM EPPS buffer (potassium phosphate buffer, 50 mM EPPS-buffer with 0.1 mM THF and 2.5 Mgl+) under 20 °C while shaking of the reaction set-up was chosen to be 600 min⁻¹.

Figure A.30: Exemplary screenshot detail from the BioCatNet web interface, showing the possibility to display a plain text summarizing the reaction conditions from the individual tables.

<table>
<thead>
<tr>
<th>Time [s]</th>
<th>Concentration [mol/l]</th>
<th>Iteration no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0011</td>
<td>1</td>
</tr>
<tr>
<td>60</td>
<td>0.06</td>
<td>1</td>
</tr>
<tr>
<td>120</td>
<td>0.0593</td>
<td>1</td>
</tr>
<tr>
<td>180</td>
<td>0.0379</td>
<td>1</td>
</tr>
<tr>
<td>360</td>
<td>0.0354</td>
<td>1</td>
</tr>
<tr>
<td>540</td>
<td>0.0329</td>
<td>1</td>
</tr>
<tr>
<td>720</td>
<td>0.0311</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure A.31: Screenshot detail from the BioCatNet web interface for a measurement of substrate and product concentrations over time (here: self-ligation from benzaldehyde to benzoin).
Table A.14: Overview of mandatory and optional input for submitting experimental data to BioCatNet. Experimental data can be experimenter-defined or measured concentrations as well as derived properties like conversion, yield or enantiomeric excess.

<table>
<thead>
<tr>
<th>topic</th>
<th>mandatory input</th>
<th>optional input</th>
</tr>
</thead>
<tbody>
<tr>
<td>reaction</td>
<td>reaction name, involved substrates and products, stoichiometry</td>
<td>SMILES code for compounds (Weininger, 1988)</td>
</tr>
<tr>
<td>reaction condition</td>
<td>initial reaction volume&lt;sup&gt;a&lt;/sup&gt;, pH, temperature, pressure, shaking frequency, further description</td>
<td></td>
</tr>
<tr>
<td>reaction buffer</td>
<td>unique name, description</td>
<td>preparation/ purification/ immobilization/ crude cell extract/ whole cell transformation, source organism (recommended), concentration or amount and volume&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>enzyme</td>
<td>exact protein sequence, sequence name, expression host, time of addition</td>
<td>preparation/ purification/ immobilization/ crude cell extract/ whole cell transformation, source organism (recommended), concentration or amount and volume&lt;sup&gt;a&lt;/sup&gt;, measurement replication number</td>
</tr>
<tr>
<td>substrate</td>
<td>compound name, time of addition or measurement</td>
<td>preparation, measurement method, concentration or amount and volume&lt;sup&gt;a&lt;/sup&gt;, measurement replication number</td>
</tr>
<tr>
<td>product</td>
<td>compound name, time of measurement</td>
<td>preparation, measurement method, concentration or amount and volume&lt;sup&gt;a&lt;/sup&gt;, measurement replication number</td>
</tr>
<tr>
<td>additive</td>
<td>compound name, time of addition</td>
<td>preparation, concentration or amount and volume&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>parameter</td>
<td>compound of reference, time of measurement, parameter, value and unit</td>
<td>parameter abbreviation, wild-type/reference sequence&lt;sup&gt;b&lt;/sup&gt;, measurement method, additional information</td>
</tr>
</tbody>
</table>

<sup>a</sup> Enzymes, substrates and additive compounds (e.g. enzyme inhibitors) are quantified by chemical amount and volume (if the initial reaction volume is specified) or by concentration.

<sup>b</sup> Some measured parameters might refer to a wild-type or reference sequence.
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

A.4.6 Supporting Information

A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

Figure A.32: Simplified scheme of the relational data model for the BioCatNet database system for experimental data. The `PARAMETER_MEASUREMENTS` and `ENZYME_FEEDS` tables link to `SEQUENCES`, which was inherited from the previously described DWARF system (Fischer et al., 2006).
A.4 BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

Figure A.33: Alignment of 30 entries in the NCBI protein database for "pyruvate decarboxylase from *Acetobacter pasteurianus*" showing multiple sequences varying both at the N- and C-termini and within the sequences (beginning on page 190). The alignment was generated using Clustal omega (Sievers *et al.*, 2011).
A database system for the integration of enzyme sequences and biocatalytic experiments

BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments

<table>
<thead>
<tr>
<th>gi</th>
<th>10385151</th>
<th>gb</th>
<th>AAM21209.1</th>
<th>AF368435_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>189974028</td>
<td>gb</td>
<td>AAM21208.1</td>
<td>AF368435_1</td>
</tr>
<tr>
<td>gi</td>
<td>529247081</td>
<td>ref</td>
<td>WP_020944449.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>915465291</td>
<td>ref</td>
<td>WP_064785391.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>1737965444</td>
<td>ref</td>
<td>WP_053646294.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>151979195</td>
<td>ref</td>
<td>WP_010008123.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>504270371</td>
<td>ref</td>
<td>WP_035366029.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>915485271</td>
<td>ref</td>
<td>WP_050818574.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>737384544</td>
<td>ref</td>
<td>WP_035366029.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>517917915</td>
<td>ref</td>
<td>WP_019088123.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>504270371</td>
<td>ref</td>
<td>WP_014457473.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>1039400602</td>
<td>ref</td>
<td>WP_064776314.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>528528981</td>
<td>emb</td>
<td>CCT59281.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>256652749</td>
<td>dbj</td>
<td>BAI18683.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>256655805</td>
<td>dbj</td>
<td>BAI21732.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>256649761</td>
<td>dbj</td>
<td>BAI15702.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>256646363</td>
<td>dbj</td>
<td>BAI09060.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>256646096</td>
<td>dbj</td>
<td>BAI05513.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>256653486</td>
<td>dbj</td>
<td>BAI04653.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>137147271</td>
<td>dbj</td>
<td>GAB31399.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>198556514</td>
<td>dbj</td>
<td>BAI09462.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>178847311</td>
<td>dbj</td>
<td>ZV911</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>178847310</td>
<td>dbj</td>
<td>ZV911</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>178847309</td>
<td>dbj</td>
<td>ZV911</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>178847308</td>
<td>dbj</td>
<td>ZV911</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>178847307</td>
<td>dbj</td>
<td>ZV911</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>178847306</td>
<td>dbj</td>
<td>ZV911</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>178847305</td>
<td>dbj</td>
<td>ZV911</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>178847304</td>
<td>dbj</td>
<td>ZV911</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>10385152</td>
<td>gb</td>
<td>AAM21209.1</td>
<td>AF368435_2</td>
</tr>
<tr>
<td>gi</td>
<td>1296524501</td>
<td>emb</td>
<td>CCB57691.1</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>157998209</td>
<td>emb</td>
<td>CA15659.1</td>
<td></td>
</tr>
</tbody>
</table>
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation


Own contributions

I curated the data provided by Zavrel et al. (2008) and Ohs et al. (2018) within the previously established BioCatNet database system. I applied the workflow for parameter estimation from Ohs et al. (2018) on both datasets for different kinetic models (see chapter 2.2.2). I contributed to the writing of the original manuscript.

A.5.1 Abstract

The estimation of kinetic parameters provides valuable insights into the function of biocatalysts and is indispensable in optimizing process conditions. Frequently, kinetic analysis relies on the Michaelis-Menten model derived from initial reaction rates at different initial substrate concentrations. However, by analysis of complete progress curves, more complex kinetic models can be identified. This case study compares two previously published experiments on benzaldehyde lyase-catalyzed self-ligation for the substrates benzaldehyde and 3,5-dimethoxybenzaldehyde to investigate (1) the effect of using different kinetic model equations on the kinetic parameter values, and (2) the effect of using models with and without enzyme inactivation on the kinetic parameter values. These analyses first highlight possible pitfalls in the interpretation of kinetic parameter estimates and second suggest a consistent strategy for data management and validation of kinetic models: First, Michaelis-Menten parameters need to be interpreted with care, complete progress curves are necessary to describe the reaction dynamics, and all experimental conditions have to be taken into consideration when interpreting parameter estimates. Second, complete progress curves should be stored together with the respective reaction conditions, to consistently annotate experimental data and avoid misinterpretation of kinetic parameters. Such data management strategy is provided by the BioCatNet database system.

A.5.2 Introduction

Modeling is a versatile tool in industrial biotechnology (Vasić-Rački et al., 2003, 2011; Almquist et al., 2014; Glynn et al., 2014). With cells as biocatalysts, mathematical models are used to understand, predict, and optimize the properties and behavior of cell factories (Tyo et al., 2010). Moreover, models are applied to suggest targets for metabolic engineering leading to increases
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

in productivity, to optimize and control biotransformation as well as fermentation processes, and to guide in the choice of cell factory when introducing a novel product (Almquist et al., 2014). With isolated enzymes as biocatalysts, enzyme kinetic models significantly increase the knowledge about the reaction systems (Vasić-Rački et al., 2003; Glynn et al., 2014), which is especially helpful if the enzyme has only recently been discovered (Baş et al., 2007). Thereby, models help to find an optimal microenvironment for the biocatalyst and optimal operating conditions to increase productivity (Vasić-Rački et al., 2003), which allows for an effective scale-up of pharmaceutical processes (Pollard and Woodley, 2007) as well as an efficient design of continuously operated enzyme reactors (Vasić-Rački et al., 2003). The main potential of enzyme kinetic models lies in identifying new process improvement strategies in silico and then using simulations to evaluate these strategies time- and resource-efficiently.

To fully exploit this potential, the enzyme kinetic models must be valid over a wide range of reaction conditions, or at least under the conditions that apply to a potential large-scale process in question (Vasić-Rački et al., 2003, 2011; Almquist et al., 2014; Wandrey, 1993; Bommarius and Riebel, 2008; Schmidt et al., 2009). More specifically, the effect of all compounds in the reaction medium on the reaction rate should be taken into account, for example, high concentrations of substrate and product as they might have a strong influence on the rate, as well as the enzyme stability under operating conditions (Vasić-Rački et al., 2011). To also detect enzyme inhibition by reaction products, kinetic analysis should favor using the complete progress curve of the reaction over only using initial velocities (Gutiérrez and Danielson, 2006). Moreover, these progress curve data can be used for detecting long-time phenomena, such as enzyme deactivation (Straathof, 2001) and for discriminating competing models (Michalik et al., 2007; Bates and Frieden, 1973).

The corresponding enzymatic experiments yield heterogeneous data sets, containing not only the observed progress curves, that is, the time course of substrate depletion or product formation, but also a description of the reaction conditions. Keeping track of the methodological details of an enzyme-catalyzed reaction is indispensable to enable the comparison of measurements. Since full-text descriptions of experimental literature cannot be read efficiently in an automated manner, databases and representative data formats comprising both experimental conditions and experimental data have been proposed to facilitate the exchange of scientific knowledge on enzymatic reactions (Gardossi et al., 2010; Buchholz et al., 2016; Tipton et al., 2014). The STRENDA consortium has suggested a comprehensive set of attributes as a benchmark for documentation of enzymatic data.\footnote{See footnote 20 on page 176.}

However, databases on literature mining of enzymatic reactions like BRENDA (Placzek et al., 2017; Schomburg et al., 2002) and SABIO-RK (Wittig et al., 2012) lack detailed information on the complete sequence of the enzyme or the description of the underlying kinetic
models, which hinders comparing and reproducing the experiments. Many kinetic analyses rely on simplified Michaelis-Menten models that are only valid under certain conditions (Dette et al., 2010; Schnell, 2014). A Michaelis-Menten model might fit the initial phase of the progress curve sufficiently well while neglecting long-term effects in the dynamics of the enzymatic reaction, such as reaction equilibrium or enzyme inactivation. Using inappropriate kinetic models hinders the mechanistic interpretation of previously published measurements, for example, in the case of regulatory interactions or enzyme inhibition (Walsh, 2014). In most cases, databases and literature rather summarize enzyme properties by listing kinetic parameters, sometimes referring to the steady-state of a reaction, instead of showing the underlying progress curve data on the dynamics of an enzymatic reaction. Thereby, it becomes impossible to analyze and compare previously published parameter estimates.

To exploit the full potential of experimental data, this study presents a workflow to measure, archive, and analyze experimental data efficiently, supported by a database system for storage and annotation of experimental data and a Matlab routine for applying different kinetic models. The results obtained from different kinetic models are compared, emphasizing the importance of a consistent data management.

A.5.3 Experimental section

Experimental data Two published datasets (Table A.15) were used to establish the workflow from measured data to kinetic model analysis. Dataset 1 covers the symmetric carboligation of 3,5-dimethoxybenzaldehyde (DMBA) molecules yielding \((R)-3,3',5,5'-\text{tetramethoxybenzoin}\) using benzaldehyde lyase (EC 4.1.2.38) from \textit{Pseudomonas fluorescens} (\textit{Pf}BAL) as biocatalyst (Zavrel et al., 2008). The assay contained 50 mM potassium phosphate (KPi) buffer as well as 0.25 mM MgCl\(_2\) and 0.25 mM thiamine diphosphate (ThDP) as cofactors. 30 \((v/v)\)% of the co-solvent dimethylsulfoxide (DMSO) was added to increase the solubility of aromatic compounds. The pH value was set to 8.5 and the ionic strength to 150 mM. The fluorescent substrate DMBA enabled the measurement of highly resolved progress curves using a fluorimeter (PerkinElmer, LS55, Waltham, MA) conducted in cuvettes at 25°C. The dataset comprises nine progress curve experiments of substrate depletion with in total 2786 data points.

Dataset 2 covers the symmetric carboligation of benzaldehyde (BA) to \((R)\)-benzoin using the same enzyme, \textit{Pf}BAL (Ohs et al., 2018), and also consists of progress curve experiments. In this case, a reaction mixture containing the enzyme \textit{Pf}BAL, buffer (50 mM triethanolamine (TEA) pH 8.5), cofactors (2.5 mM MgSO\(_4\), 0.5 mM ThDP), co-solvent (30 \((v/v)\)% DMSO) and substrate was distributed to 1800 \(\mu\text{L}\) HPLC vials after mixing. The vials were immediately closed to prevent evaporation of benzaldehyde, then located into a heat shaker (preheated to 30°C) and shaken at 300 rpm. At specific times, 600 \(\mu\text{L}\) samples were taken from one vial,
and the reaction was stopped by adding 300 µL trichloroacetic acid. Subsequently, the sample was diluted and analyzed using high performance liquid chromatography (HPLC). The dataset comprises 13 progress curve experiments of substrate depletion and product formation with in total 374 data points. As a test for data quality, the mass balance was checked for each experiment.

The progress curve data from datasets 1 and 2 are available as supplementary files (40 and 41, respectively).

**Workflow for kinetic parameter estimation** The workflow of data management and data analysis combines the consistent data model of the BioCatNet database system with advanced routines for kinetic parameter estimation.

Two datasets of enzyme-catalyzed self-ligation with either 3,5-dimethoxybenzaldehyde (DMBA) or benzaldehyde (BA) as substrate were analyzed by four different kinetic models, and the parameter estimation was performed in three subsequent steps of Monte-Carlo, algebraic, and dynamic parameter estimation to gradually increase the accuracy of the estimation while saving computation time (Figure A.34).

(1) **Data management**

The BioCatNet database system stores data on protein sequences, protein structures and biocatalytic experimental data, preferably progress curves of substrate depletion and/or product formation, providing a starting point for the comparative analysis of experimental data and a link to sequences and structural data for the applied enzymes. Experimental data are described by several attributes, including information on the initial concentrations, the reaction buffer, and additives, for example, co-solvents. In addition, the full protein sequence of the applied enzyme is linked to the experiment. The two datasets for self-ligation of BA and DMBA were uploaded as experiment sets into the Thiamine diphosphate-dependent Enzymes Engineering Database.

(2) **Matlab tool for kinetic analysis of enzyme reactions**

The kinetic analysis of the two datasets is conducted using a Matlab-based tool for kinetic analysis of carboligation reactions (Ohs et al., 2018). This tool can be expanded by additional kinetic models to allow for analysis of further enzyme-catalyzed reactions, see below. Here, the kinetic analysis starts with choosing a reaction model from a pool of implemented models according to the reaction mechanism. In a second step, boundaries for the kinetic parameters are specified. Afterwards, the experimental data sets containing the progress curve data are read in. The tool then interpolates the reaction rates from these data. The subsequent parameter

41 SuppTab-S2.xlsx, accessed on June 14, 2018.
42 See footnote 15 on page 174.
estimation comprises three steps, first, a parameter estimation based on reaction rates by Monte-Carlo simulation with a user-defined sample size of random parameter values, second, an algebraic parameter estimation also based on the reaction rates, and third, a dynamic parameter estimation based on the original progress curve data. The overall parameter estimation routine comprising all three steps was repeated five times. Finally, the quality of the model fit is returned, together with the relative standard deviations of the parameter estimates, derived from the covariance matrix. Details on the parameter estimation routine are given in the supplementary material. The parameter set resulting in the lowest residual sum of squares (RSS) is reported as result of the parameter estimation. Thereby, the tool can easily be used to compare different models. The Matlab code is available for download at 43.

Model equation for enzyme kinetics In the following, the term model refers to a mathematical model, described by an equation system. One underlying reaction mechanism may be represented by several mathematical models, that is different model equations, based on different assumptions each. Model 1 is the irreversible macrokinetic Michaelis-Menten model, with adapted stoichiometry for the symmetric carboligation (Figure A.35) and which does not consider enzyme inactivation:

\[
\frac{dc_A}{dt} = -2 \frac{dc_P}{dt} = \frac{k_{cat} c_{Et} c_A}{K_{mA} + c_A} \tag{A.1}
\]

where \(t\) is time, \(c_A\) is the substrate concentration, \(c_P\) is the product concentration, \(c_{Et}\) is the total enzyme concentration, and the two model parameters \(k_{cat}\) are the rate constant for the forward reaction and \(K_{mA}\) the Michaelis-Menten constant, respectively. The macrokinetic Michaelis-Menten parameter lumps several elementary rate parameters into one effective parameter, which describes the substrate binding affinity.

Model 2 uses macrokinetic rate expressions, including reversibility of the reaction (Zavrel et al., 2008) and does not consider enzyme inactivation. It takes all elementary reaction steps of the ordered bi-uni mechanism (Figure A.35) (Zavrel et al., 2008; Ohs et al., 2018) into account:

\[
\frac{dc_p}{dt} = -\frac{1}{2} \frac{dc_A}{dt} = \frac{N_1}{D_1} \tag{A.2}
\]

\[
N_1 = \frac{k_{cat}}{K_i A K_mB} \left( c_A^2 - \frac{c_P}{K_{eq}} \right) c_{Et} \tag{A.3}
\]

\[
D_1 = 1 + \frac{c_A}{K_i A} + \frac{K_{mA} c_A}{K_i A K_mB} + \frac{c_A^2}{K_i A K_mB} + \frac{c_P}{K_mB} + \frac{c_A c_P}{K_mC_iB} \tag{A.4}
\]

A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

with the Haldane relations

\[ K_{iB} = \frac{K_mB K_iA}{K_mA \left(1 - \left(\frac{K_mA}{K_iA} - 1\right) \frac{K_mP}{K_eq K_mB K_iA}\right)} \] (A.5)

\[ K_mP = \frac{K_mB (K_mB - K_mA)^2 K_eq}{2K_mA} \] (A.6)

\[ K_iA = K_mB - K_mA \] (A.7)

\[ K_eq = \frac{c_P}{c_A} \] (A.8)

with the four model parameters \( k_{\text{catf}} \) as the rate constant for the forward reaction, \( K_eq \) as the equilibrium constant, and \( K_mA \) as well as \( K_mB \) as the Michaelis-Menten constants for the first binding (donor) and the second binding substrate (acceptor), respectively. The parameters \( K_mP, K_iA, \) and \( K_iB \) are the Michaelis-Menten constants for the product (reverse reaction), the inhibition constants of the first binding substrate and the second binding substrate respectively, which are derived from the model parameters by the Haldane equations (Zavrel et al., 2008). Here, the model assumes that the kinetic parameters for substrate binding and release are identical \((k_2 = k_1, k_{-2} = k_{-1})\). The last two Haldane relations (Equation A.7 and A.8) result from this assumption. In this study, model 2 is called simplified macrokinetic self-ligation model.

**Model 3** is a microkinetic model, again without considering enzyme inactivation. It is a different mathematical formulation of the same reaction mechanism underlying also model 2, namely in elementary rate parameters instead of effective macroscopic rate parameters (Zavrel et al., 2008). The model equations for model 3 read as:

\[ \frac{dc_P}{dt} = -\frac{1}{2} \frac{dc_A}{dt} = \frac{N_2}{D_2} \] (A.10)

\[ N_2 = (k_1^2 k_3 c_A^2 - k_{-1}^2 k_{-3} c_P) c_{Ei} \] (A.11)

\[ D_2 = k_{-1}^2 + k_{-1} k_3 + k_1^2 c_A^2 + k_1 k_{-3} c_A c_P + (2 k_1 k_3 + k_1 k_{-1}) c_A + 2 k_{-1} k_{-3} c_P \] (A.12)

where \( k_i \) are the rate parameters for the corresponding elementary reactions, in the following named microkinetic parameters. Model 3 is called simplified microkinetic self-ligation model (Zavrel et al., 2008; Ohs et al., 2018), as it also assumes that the two kinetic parameters for substrate binding and release are identical \((k_2 = k_1, k_{-2} = k_{-1})\), leading to four independent model parameters \( k_1, k_{-1}, k_3, \) and \( k_{-3} \). The macro- and microkinetic models are equivalent, since both models represent the same first principles of the bi-uni reaction mechanism. There-
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Therefore, the kinetic parameters can be converted into each other. The relations to calculate the microkinetic parameters from the macrokinetic parameters are:

\[
\begin{align*}
  k_1 &= k_2 = \frac{k_{\text{cat}f}}{K_{mA}} \\
  k_{-1} &= k_{-2} = \frac{k_{\text{cat}f}(K_mB - K_mA)}{K_mA} \\
  k_3 &= k_{\text{cat}f} \\
  k_{-3} &= \frac{k_{\text{cat}f}}{K_{eq}(K_mB - K_mA)}
\end{align*}
\]  

(A.13) 

(A.14) 

(A.15) 

(A.16)

and the ones to calculate the macrokinetic parameters from the microkinetic ones:

\[
\begin{align*}
  k_{\text{cat}f} &= k_3 \\
  K_{mA} &= \frac{k_3}{k_1} \\
  K_mB &= \frac{k_{-1} + k_3}{k_1} \\
  K_{eq} &= \frac{k_1^2 k_3}{k_{-1} k_{-3}}
\end{align*}
\]

(A.17) 

(A.18) 

(A.19) 

(A.20)

While \( k_{\text{cat}f}, K_mA \) and \( K_mB \) are enzyme-specific constants, \( K_{eq} \) is enzyme-independent and only depends on the chemical equilibrium of substrate and product. This fact implies that \( k_1, k_3, k_{-1} \) and \( k_{-3} \) are linked by an external thermodynamic constraint.

Model 4 is a microkinetic model which considers substrate-dependent enzyme inactivation (Ohs et al., 2018) by adding an inactivation term to model 3. The equation for the inactivation reads as:

\[
\frac{dc_E}{dt} = -k_{inS} \cdot c_A \cdot c_E
\]

(A.21)

where \( k_{inS} \) is the inactivation parameter.

A.5.4 Results

Michaelis-Menten model 1 is only applicable for the initial part of the progress curve. The Michaelis-Menten model 1 is the standard choice for evaluating enzyme kinetics from initial rate data. Indeed, since the carboligation reactions are reversible, the irreversible Michaelis-Menten model 1 cannot sufficiently fit the complete progress curve data (Figures A.36 and A.37). As a result, the estimated macroscopic kinetic parameters \( K mA \) and \( k_{\text{cat}f} \) show high uncertainty and cannot be identified reliably (Table A.18). Therefore,
the Michaelis-Menten model 1 was applied to the first six minutes of the progress curve, corresponding to typical initial rate measurements with presumably low product concentrations. For both substrates, the shortened progress curve subset resulted in an improved model fit (Figures A.38 and A.39) and in macroscopic kinetic parameters $K_{mA}$ and $k_{catf}$ with remarkably reduced uncertainty in comparison to the full progress curves (Tables A.16 and A.17). The evaluation of the first six minutes of the reaction progress covers a suitable range of substrate concentrations between 1.5 and 3 mM for DMBA and between 30 and 75 mM for BA (Table A.15), below and above the respective $K_{mA}$-values of 2.6 mM and 55 mM of model 1 (Tables A.16 and A.17). While the $K_{mA}$ values differ by an order of magnitude, the $k_{catf}$ values are very similar for both substrates.

Michaelis-Menten model 1 provides better parameter quality than the macrokinetic model 2 for the initial reaction phase The resulting kinetic parameters of the Michaelis-Menten model 1 covering the shortened progress curves were compared to the previously published macrokinetic bi-uni model 2 covering the full progress curves (Tables A.16 and A.17). Model 2 for the bi-uni reaction mechanism has been applied previously to the symmetric carboligation of DMBA (Zavrel et al., 2008). While for both substrates the $k_{catf}$ values were similar to model 1, the $K_m$ values were consistent with model 1 for BA, but differed by up to three orders of magnitude for DMBA. The $K_m$-values of the bi-uni model 2 showed significantly larger relative standard deviations of more than 100%, meaning that both $K_m$ parameters were unidentifiable, in contrast to the $K_m$ returned by model 1 for the initial progress curves.

Microkinetic model 3 results in more certain parameter estimates than macrokinetic model 2 As expected, the macrokinetic and the microkinetic model for the reversible bi-uni reaction without inactivation had similar $RSS/n$ for both substrates, indicating a similar overall fit quality (Figures A.40 to A.43). For DMBA, the $k_{catf}$ (equivalent to $k_3$), and $K_{eq}$ values were found identical, but the $K_{mA}$ and $K_{mB}$ as well as the other microkinetic parameters differed by an order of magnitude for models 2 and 3 (Table A.16). Noteworthy, most relative standard deviations for the parameter estimates were remarkably smaller for the microkinetic model 3 than for the previously published macrokinetic model 2, although both models describe the same reaction mechanism. For BA, all kinetic parameters differed by less than a factor of two. The low $K_m$ values for DMBA and the high $K_{eq}$ for BA give first hints that kinetic models 2 and 3 are unsuitable, in contrast to the previously published kinetic analysis for DMBA (Zavrel et al., 2008).
Microkinetic model 3 is outperformed by microkinetic model 4 with substrate-dependent inactivation Since the symmetric carboligation of BA has been studied with respect to substrate-dependent enzyme inactivation (Ohs et al., 2018), it seemed promising to test model 4 on the DMBA dataset, too. For both substrates, the $RSS/n$ and the relative standard deviations of the kinetic parameters were remarkably lower for microkinetic model 4 with substrate-dependent inactivation, indicating a better model quality and improved identifiability of the parameter estimates (Figures A.44 and A.45). Unexpectedly, the equilibrium constant differed by a factor of twenty between both substrates. The catalytic activity of PfBAL towards both substrates differed by less than a factor of two, whereas $K_{mA}$ towards DMBA was more than two orders of magnitude lower than $K_{mA}$ towards BA (Tables A.16 and A.17). The differences in $K_{mA}$ and $K_{mB}$ between the substrates correlate to the two orders of magnitude difference in $k_1$ and $k_{-3}$ between the substrates, whereas $k_{-1}$ and $k_3$ for the two substrates were similar. The inactivation constant $k_{inS}$ for BA was one order of magnitude smaller than the one of DMBA.

A.5.5 Discussion

Model and parameter quality hint at substrate-dependent inactivation mechanism

The model fit and the parameter quality of the microkinetic model 4 with substrate-dependent inactivation clearly support the mechanism of substrate-dependent inactivation for both substrates as observed previously (Leksawasdi et al., 2004). Michaelis-Menten model 1, which has a similar fit quality and similar $k_{catf}$ values, cannot capture inactivation since it neglects the later parts of the progress curves. It is interesting, that the kinetic models 2 and 3 without inactivation term misinterpret the effect of inactivation on the reaction rate in two different ways: for BA as a shift in reaction equilibrium ($K_{eq}$), but for DMBA as a shift in substrate affinity ($K_{mA}$). This can be explained by differences in the experimental data (Figures A.42 to A.45): While the DMBA progress curves reached reaction equilibrium, and thus led to reliable estimates of $K_{eq}$, the BA progress curves do not fully reach the equilibrium, and leave more freedom to the estimation of $K_{eq}$. The 10-fold higher rate constant for substrate-dependent enzyme inactivation for DMBA correlates with its higher binding affinity, which suggests that inactivation might be caused by substrate binding near to the active site. However, at respective substrate concentrations of 3 mM DMBA and 50 mM BA, half times $\tau_{1/2}$ of PfBAL inactivation ($\tau_{1/2} = (k_{inS} \cdot c_A)^{-1}$) are 40 and 30 min, respectively, which precludes mere binding effects.

Interpretation of the Michaelis-Menten parameters

In the microkinetic model 4 with substrate-dependent inactivation, $K_{mA}$ is smaller than $K_{mB}$ by a factor 1.5 to 3 since the probability of substrate binding is higher upon the first binding step as reflected by Equations A.18 and A.19. For BA, the $K_{mA}$ value of the Michaelis-Menten model 1 (55 mM)
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

is consistent with the $K_{mA}$ and $K_{mB}$ values (20 and 32 mM, respectively) of the microkinetic model 4 with inactivation, because model 4 describes the substrate depletion as a subsequent two-step-mechanism, each with separate binding affinities $K_{mA}$ and $K_{mB}$, whereas model 1 combines both binding events into a single substrate affinity, which is accordingly higher than the individual binding affinities. For DMBA, the trend is similar. The differences between the $K_{mA}$ and $K_{mB}$ values between model 1 and model 4, however, is not caused by product inhibition at low conversion (Pleiss, 2017), because the Michaelis-Menten parameters $K_{mP}$ (0.12 mM for 3,3',5,5'-tetramethoxybenzoin and 23 mM for benzoin) as derived from model 4 (Equation A.6) were similar to the respective $K_{mB}$ values (0.2 mM and 32 mM, respectively).

Solvent effects and substrate solubility The $K_{mA}$ and $K_{mB}$ values for BA of 20 and 32 mM at 30 % DMSO are in agreement with previously published data showing that $K_{mA}$ and $K_{mB}$ increase with increasing DMSO co-solvent content from 0.28 and 0.39 mM without DMSO to 1.96 and 6.13 mM at 10 % DMSO, respectively (Kokova et al., 2009). The decrease of the apparent affinity to the hydrophobic active site by two orders of magnitude at 30 % DMSO is a consequence of two factors, one being the decrease of the activity coefficients of the substrates and products (Pleiss, 2017) and the other being the competitive inhibition of PfBAL by DMSO (Gerhards et al., 2012). Although BA is considered to be the “natural” substrate of PfBAL, its $k_{cat}$ value was similar to the “unnatural” DMBA, but its $K_{mB}$ was two orders of magnitude higher than $K_{mB}$ of DMBA. Because the $K_{mB}$ value results from two main phenomena, the enzyme-substrate interaction and the substrate-solvent interaction, the lower $K_{mB}$ value of DMBA can be partially explained by its sixfold lower estimated aqueous solubility resulting in sixfold higher activity coefficient of DMBA. A corresponding separation of the substrate-solvent effects from the substrate-enzyme effects has been accomplished by expressing the kinetic parameters in terms of thermodynamic activities (Grosch et al., 2017). In general, it is advisable to annotate reaction conditions consistently, including the composition of the reaction buffer and co-solvents, since this information is a prerequisite to capture solvent effects.

Lessons learned Although the reaction system of this case study is not overly complex, researchers unfamiliar with the concepts of kinetic modeling might encounter pitfalls when trying to reliably estimate kinetic parameters. To avoid misinterpretations and to enhance reproducibility of results, several points have to be noted:

1. Be careful with the interpretation of Michaelis-Menten parameters. The basic phenomena of the substrates’ affinity can be interpreted using the Michaelis-Menten model 1, but only the mechanistic microkinetic model 4 including substrate-dependent inactivation can be interpreted in detail by calculating the corresponding microkinetic parameters for $K_{mA}$.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

and $K_{mB}$. Only then, the elementary reaction rates allow for identifying rate-limiting steps.

2. Use progress curves for kinetic analysis of the reaction dynamics. Our findings show that only the initial portion, but not the full progress curves are well fit by the Michaelis-Menten model, and confirm Hill et al. (1977) analysis that literature does not widely support the Michaelis-Menten model. As experienced kinetic modelers know, only complete progress curves reveal phenomena like reversibility, inhibition or inactivation, which are not yet observable during the initial phase of the reaction when product formation is still low. The better model fit of the Michaelis-Menten model to the initial fraction in comparison to the full progress curves does not guarantee that this model sufficiently describes the underlying mechanism (Stroberg and Schnell, 2016). When using progress curves for estimating kinetic parameters, the biochemist avoids the information loss of the rate regression from the concentration over time data and is rewarded by more precise information about the full enzymatic reaction, in contrast to investigations of the initial reaction phase or steady-state kinetics (Schnell and Maini, 2003).

3. Consider all relevant phenomena in mechanistic modelling. Only when substrate-dependent inactivation was considered, the mechanistic modelling led to significantly better results than the Michaelis-Menten model and to correct interpretation of the occurring phenomena. In other words, mechanistic modelling needs to be sufficiently detailed and describe not only the reaction mechanism and the accompanying reaction equilibria, but also reflect enzyme inactivation (Rachinskiy et al., 2009) and for example, substrate instability by suitable expressions. If additional influence factors such as co-solvent concentration, temperature, and/or pH-value are varied, thermodynamic activities may replace the concentrations, and additional terms may complement the microkinetic models. The model might be further refined by dropping the assumptions $k_1 = k_2$ and $k_{-1} = k_{-2}$ (Zavrel et al., 2008) The selection of a suitable kinetic model did not only influence the mechanistic interpretation of reactions in biochemistry and in pharmaceutical studies (Walsh, 2014), but has also been pivotal for determining meaningful cytochrome P450 enzyme kinetics (Tracy, 2006; Kramer and Tracy, 2008).

4. Manage your experimental data well. Two prerequisites have to be fulfilled to re-analyze progress curve data from different data sources with new models: First, efficient, stepwise parameter estimation reduces the effort for progress curve integration (Ohs et al., 2018). The Monte Carlo methods used therein have also stood the test in more complex kinetic models (Srinivasan et al., 2015). Second, only the storage of progress curve raw data on a publicly available platform like the BioCatNet database system allows for evaluating
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

alternative kinetic models. In structural biology, the problem of misinterpretation of protein structures has been recognized many years ago, and nowadays, deposited original data enable alternative approaches to structure refinement (Jones et al., 1996; Kleywegt and Jones, 1998). Similarly, the BioCatNet concept of providing original data will be helpful in the management of higher amounts of data, such as combinations of enzyme variants with progress curve data measured under various initial conditions.

5. Annotate experimental conditions consistently. Likewise, biocatalytic data should be consistently stored together with the respective reaction conditions as an essential prerequisite for a reproducible kinetic analysis (Gardossi et al., 2010). Similar critical requirements for managing enzymatic data were formulated by the STRENDATA consortium (Tipton et al., 2014) and database projects such as SABIO-RK (Wittig et al., 2012) and BioCatNet, with the latter also comprising the original progress curve data of a biocatalytic experiment (Buchholz et al., 2016). Finally, kinetic analyses need to be linked not only to the experimental conditions, but also to the corresponding kinetic models and to the protein sequence or structure data. The integration of experimental biocatalytic data with information on the respective enzyme sequence will support the systematic study of sequence-structure-function relationships and, based on the knowledge obtained, the design of improved enzyme variants and bioprocesses.

Acknowledgement

The authors gratefully acknowledge funding through the German Research Foundation (DFG) in the framework of the Research Group ”Diversity in Asymmetric Thiamine Catalysis” (FOR 1296) and by grant PL145/16-1.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.34: Simplified scheme of the workflow that combines experimental data retrieved from the BioCatNet database with routines for the estimation of kinetic parameters. In each repetition of the kinetic parameter estimation, three subsequent steps of parameter estimation are performed as a trade-off between accuracy and performance (Ohs et al., 2018), using previously interpolated reaction rates (Monte-Carlo and algebraic parameter estimation) and finally using progress curves of concentrations over time (dynamic parameter estimation). Quality indicators comprise standard deviations of kinetic parameters and the residual sum of squares ($\text{RSS}$) with respect to the full data set.

![Figure A.34: Simplified scheme of the workflow that combines experimental data retrieved from the BioCatNet database with routines for the estimation of kinetic parameters.](image)

Figure A.35: Microkinetic representation of the Michaelis-Menten model for symmetric carboligation (model 1) with substrate (A), product (P), free enzyme (E), and enzyme-substrate/product-complexes (EA, EAA, EP). For the microkinetic model of the ordered bi-uni mechanism (models 2 to 4), it is assumed that for the first and second binding substrate (A and B) $k_1 = k_2$ and $k_{-1} = k_{-2}$ (Zavrel et al., 2008; Ohs et al., 2018).

![Figure A.35: Microkinetic representation of the Michaelis-Menten model for symmetric carboligation (model 1) with substrate (A), product (P), free enzyme (E), and enzyme-substrate/product-complexes (EA, EAA, EP). For the microkinetic model of the ordered bi-uni mechanism (models 2 to 4), it is assumed that for the first and second binding substrate (A and B) $k_1 = k_2$ and $k_{-1} = k_{-2}$ (Zavrel et al., 2008; Ohs et al., 2018).](image)
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Table A.15: Initial experimental conditions.

<table>
<thead>
<tr>
<th>Dataset 1 (DMBA)</th>
<th>$c_A(0)$ [mM]</th>
<th>$c_P(0)$ [mM]</th>
<th>$c_{Et}$ [mM]</th>
<th>duration [min]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>8.33 $\cdot$ $10^{-5}$</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2.75</td>
<td>8.33 $\cdot$ $10^{-5}$</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>2.5</td>
<td>8.33 $\cdot$ $10^{-5}$</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>2.25</td>
<td>9.20 $\cdot$ $10^{-5}$</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>1.00 $\cdot$ $10^{-4}$</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>1.5</td>
<td>8.33 $\cdot$ $10^{-5}$</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>4.17 $\cdot$ $10^{-5}$</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>2.75</td>
<td>4.17 $\cdot$ $10^{-5}$</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2.5</td>
<td>4.17 $\cdot$ $10^{-5}$</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset 2 (BA)</th>
<th>$c_A(0)$ [mM]</th>
<th>$c_P(0)$ [mM]</th>
<th>$c_{Et}$ [mM]</th>
<th>duration [min]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>72.5</td>
<td>4.75 $\cdot$ $10^{-4}$</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>36.25</td>
<td>9.33 $\cdot$ $10^{-4}$</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>36.25</td>
<td>4.75 $\cdot$ $10^{-4}$</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>72.5</td>
<td>2.04 $\cdot$ $10^{-4}$</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>72.5</td>
<td>4.41 $\cdot$ $10^{-4}$</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>30</td>
<td>3.39 $\cdot$ $10^{-5}$</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>75</td>
<td>6.79 $\cdot$ $10^{-5}$</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>6.79 $\cdot$ $10^{-5}$</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>35</td>
<td>3.22 $\cdot$ $10^{-4}$</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>75</td>
<td>3.34 $\cdot$ $10^{-4}$</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>30</td>
<td>8.36 $\cdot$ $10^{-5}$</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>72.5</td>
<td>4.75 $\cdot$ $10^{-4}$</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>75</td>
<td>1.25 $\cdot$ $10^{-4}$</td>
<td>105</td>
<td></td>
</tr>
</tbody>
</table>

1 Dataset 1 (Zavrel et al., 2008) comprises 9 progress curves (A-I).
2 Dataset 2 (Ohs et al., 2018) comprises 13 progress curves (A-M).
3 Initial concentrations of substrate $A$ and product $P$.
4 Enzyme concentration.
Table A.16: Estimated kinetic parameters and the residual sum of squares (RSS) over sample size \( n \) for the benzaldehyde lyase-catalyzed self-ligation of 3,5-dimethoxybenzaldehyde and four kinetic models. Estimated parameter values are given in bold, whereas parameter values obtained from Equations A.10 to A.20 are set in regular font. The relative standard deviations (in %) are given in parentheses. Relative standard deviation values < 100 % are marked with * to indicate the higher reliability of the respective kinetic parameter values.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_1 ) mM(^{-1} ) min(^{-1} )</td>
<td>1.6 ( \cdot ) 10(^7 )</td>
<td>9.3 ( \cdot ) 10(^5 )</td>
<td>4.5 ( \cdot ) 10(^4 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(501)</td>
<td>(108)</td>
<td>(14)*</td>
<td></td>
</tr>
<tr>
<td>( k_{-1} ) min(^{-1} )</td>
<td>6.4 ( \cdot ) 10(^4 )</td>
<td>1.6 ( \cdot ) 10(^4 )</td>
<td>4.3 ( \cdot ) 10(^3 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(590)</td>
<td>(53)*</td>
<td>(6)*</td>
<td></td>
</tr>
<tr>
<td>( k_3 ) min(^{-1} )</td>
<td>2.7 ( \cdot ) 10(^3 )</td>
<td>2.7 ( \cdot ) 10(^3 )</td>
<td>2.8 ( \cdot ) 10(^3 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1)*</td>
<td>(1)*</td>
<td>(1)*</td>
<td></td>
</tr>
<tr>
<td>( k_{-3} ) mM(^{-1} ) min(^{-1} )</td>
<td>4.2 ( \cdot ) 10(^7 )</td>
<td>2.4 ( \cdot ) 10(^6 )</td>
<td>7.7 ( \cdot ) 10(^4 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(547)</td>
<td>(110)</td>
<td>(16)*</td>
<td></td>
</tr>
<tr>
<td>( k_{inS} ) mM(^{-1} ) min(^{-1} )</td>
<td>4.4 ( \cdot ) 10(^3 )</td>
<td>2.7 ( \cdot ) 10(^3 )</td>
<td>2.7 ( \cdot ) 10(^3 )</td>
<td>2.8 ( \cdot ) 10(^3 )</td>
</tr>
<tr>
<td></td>
<td>(5)*</td>
<td>(1)*</td>
<td>(1)*</td>
<td></td>
</tr>
<tr>
<td>( k_{catf} ) min(^{-1} )</td>
<td>2.6</td>
<td>1.7 ( \cdot ) 10(^{-4} )</td>
<td>2.9 ( \cdot ) 10(^{-3} )</td>
<td>6.2 ( \cdot ) 10(^{-2} )</td>
</tr>
<tr>
<td></td>
<td>(9)*</td>
<td>(1)*</td>
<td>(1)*</td>
<td></td>
</tr>
<tr>
<td>( K_{mA} ) mM</td>
<td>4.2 ( \cdot ) 10(^{-3} )</td>
<td>2.0 ( \cdot ) 10(^{-2} )</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(262)</td>
<td>(117)</td>
<td>(15)*</td>
<td></td>
</tr>
<tr>
<td>( K_{mB} ) mM</td>
<td>-</td>
<td>3.9</td>
<td>3.9</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>(1)*</td>
<td>(364)</td>
<td>(35)*</td>
<td></td>
</tr>
<tr>
<td>( K_{eq} ) mM(^{-1} )</td>
<td>6.7 ( \cdot ) 10(^{-3} )</td>
<td>1.4 ( \cdot ) 10(^{-3} )</td>
<td>1.4 ( \cdot ) 10(^{-3} )</td>
<td>1.1 ( \cdot ) 10(^{-3} )</td>
</tr>
</tbody>
</table>

1 Estimated parameters, their relative standard deviation (%), and the residual sum of squares (RSS) over sample size \( n \) were applied to nine experiments containing \( n = 2786 \) samples. Relative standard deviations of independent parameters were determined using the covariance matrix. Error propagation was used to determine relative standard deviations of derived parameter values. Compare supplementary material for details.

2 Model 1: Michaelis-Menten, model 2: macrokinetic model without inactivation, model 3: microkinetic model without inactivation, model 4: microkinetic model including substrate-dependent inactivation

3 Model 1 was applied for a maximum reaction time of 6 min, resulting in 549 samples. Parameters of model 1 for the full progress curves are listed in Table A.18.

4 Values for \( k_1 \), \( k_{-1} \), \( k_3 \) and \( k_{-3} \) were calculated from the macrokinetic parameters according to Equations A.13 to A.16 including propagation of uncertainty.

5 Values for \( k_{catf} \), \( K_{mA} \), \( K_{mB} \) and \( K_{eq} \) were calculated from the microkinetic parameters according to Equations A.17 to A.20 including propagation of uncertainty.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Table A.17: Estimated kinetic parameters and the residual sum of squares ($RSS$) over sample size $n$ for the benzaldehyde lyase-catalyzed self-ligation of benzaldehyde and four kinetic models $^6$. Footnotes and descriptions are equivalent to Table A.16.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Model 1 $^7$</th>
<th>Model 2 $^4$</th>
<th>Model 3 $^5$</th>
<th>Model 4 $^5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$ mM$^{-1}$ min$^{-1}$</td>
<td>-</td>
<td>135</td>
<td>134</td>
<td>245</td>
</tr>
<tr>
<td>$k_{-1}$ min$^{-1}$</td>
<td></td>
<td>(142)</td>
<td>(77)$^*$</td>
<td>(22)$^*$</td>
</tr>
<tr>
<td>$k_3$ min$^{-1}$</td>
<td>-</td>
<td>3.6</td>
<td>2.2</td>
<td>$3.1 \cdot 10^3$</td>
</tr>
<tr>
<td>$k_{-3}$ mM$^{-1}$ min$^{-1}$</td>
<td></td>
<td>(1.4 $\cdot 10^5$)</td>
<td>(3 $\cdot 10^4$)</td>
<td>(37)$^*$</td>
</tr>
<tr>
<td>$k_{inS}$ mM$^{-1}$ min$^{-1}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>$7.4 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>$k_{catf}$ min$^{-1}$</td>
<td>5.0 $\cdot 10^3$</td>
<td>3.0 $\cdot 10^3$</td>
<td>3.1 $\cdot 10^3$</td>
<td>4.9 $\cdot 10^3$</td>
</tr>
<tr>
<td></td>
<td>(15)$^*$</td>
<td>(55)$^*$</td>
<td>(54)$^*$</td>
<td>(12)$^*$</td>
</tr>
<tr>
<td>$K_{mA}$ mM</td>
<td>55</td>
<td>22</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>(28)$^*$</td>
<td>(131)</td>
<td>(94)$^*$</td>
<td>(25)$^*$</td>
</tr>
<tr>
<td>$K_{mB}$ mM</td>
<td>-</td>
<td>22</td>
<td>23</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(113)</td>
<td>(97)$^*$</td>
<td>(27)$^*$</td>
</tr>
<tr>
<td>$K_{eq}$ mM$^{-1}$</td>
<td>3.0 $\cdot 10^3$</td>
<td>8.6 $\cdot 10^3$</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3.6 $\cdot 10^4$)</td>
<td>(6.0 $\cdot 10^4$)</td>
<td>(94)$^*$</td>
</tr>
<tr>
<td>$RSS_n$ mM$^2$</td>
<td>1.4</td>
<td>15</td>
<td>15</td>
<td>2.7</td>
</tr>
</tbody>
</table>

$^6$ Estimated parameters, their relative standard deviation (%), and the residual sum of squares ($RSS$) over sample size $n$ were applied to thirteen experiments containing $n = 374$ samples.

$^7$ Model 1 was applied for a maximum reaction time of 6 min, resulting in 110 samples. Parameters of model 1 for the full progress curves are listed in Table A.18.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

A.5.6 Supporting Information


**Parameter estimation procedure**

Each iteration of the parameter estimation routine comprised three steps:

1. Monte-Carlo-based parameter estimation,
2. algebraic and
3. dynamic parameter estimation:

- The Monte-Carlo simulation in step (1) generates a user-defined number (here, 100,000) of random parameter sets that serve to calculate reaction rates using the kinetic models. The lowest $RSS$ between these calculated rates and the rates interpolated from the experimental data defines the best parameter set, which is used as starting point for the subsequent step (2).

- The algebraic parameter estimation in step (2) uses the lsqnonlin solver from Matlab and the interpolated reaction rates. *Lsqnonlin* is a solver suitable to optimize least-squares curve fitting problems and uses the residuals as objective function, i.e. it minimizes the distances between input data and fitted data. A maximum of 2000 iterations and the specific ”Jacobian” option for better convergence were set as solver options. The parameter estimates from step (2) were used as input for dynamic parameter estimation routine in step (3)

- In the dynamic parameter estimation in step (3), the ordinary differential equation of the respective kinetic model was solved to finally fit the model against the actual progress curve data, i.e. concentrations over time. Step (3) used the *lsqnonlin* solver with the same settings as for the previous step (2)

The parameter estimation routine was repeated five times. Subsequently, the independent estimates were compared, and the parameters of the model fit with the lowest residual sum of squares were chosen as the final parameter estimates.

All parameter estimations were performed in Matlab (The Mathworks, Natick, MA, USA, version R2016a) with upper and lower bounds for the parameter estimates set to $10^{-3}$ and $10^{3}$, respectively.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Errors of the parameter estimates
The errors of the parameter estimates were determined by analyzing the covariance matrix of the estimates. The covariance matrix $V$ was approximated as follows:

$$V(\hat{\theta}) = (Q^TQ)^{-1} s^2 = (Q^TQ)^{-1} \frac{RSS(\hat{\theta})}{n-p},$$  \hspace{1cm} (A.22)

where $\hat{\theta}$ is the best available parameters estimate, $Q$ is the (Jacobian) sensitivity matrix, $s^2$ is the estimate of the unknown measurement variance (assuming uncorrelated and constant measurement errors), $RSS(\hat{\theta})$ is the residual sum of squares, $n$ the total number of measurements, and $p$ the number of parameters (Ohs et al., 2018). The standard deviation $sd(\hat{\theta})$ and the relative standard deviation $RSD(\hat{\theta})$ were calculated as

$$sd(\hat{\theta}) = (v_{ii})^{\frac{1}{2}}$$  \hspace{1cm} (A.23)

$$RSD(\hat{\theta}) = \frac{sd(\hat{\theta}_i)}{\hat{\theta}_i}$$  \hspace{1cm} (A.24)

where the variance $v_{ii}$ is the $i$th diagonal entry of the covariance matrix $V$.

The relative standard deviations of the corresponding macrokinetic parameters of the microkinetic model, or vice versa, were calculated using the error propagation equation (Zavrel et al., 2008)

$$\sigma_K = \sqrt{\sum_{i=1}^{n} \left( \frac{\partial K}{\partial p_i} \cdot \sigma_i \right)^2}$$  \hspace{1cm} (A.25)

where $\sigma_K$ is the standard deviation of the dependent parameter $K$ and $n$ is the number of independent parameters $p_i$ with respective standard deviations $\sigma_i$. 

211
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.36: Fits of the two-substrate Michaelis-Menten model (model 1, dotted lines) to the full progress curve of dimethoxy-benzaldehyde self-ligation reactions (red points (Zavrel et al., 2008)). The reaction was performed in 50 mM KPi buffer with 0.25 mm MgCl₂, 0.25 mM ThDP and 30 % DMSO at pH 8.5 and 25 °C. The initial concentrations of substrate and enzyme are given in Table A.15. The respective parameter estimates are given in Table A.18.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.37: Fits for the benzaldehyde self-ligation (Ohs et al., 2018) and the two-substrate Michaelis-Menten model (model 1). Fitted data are indicated as lines, experimental measurement data as points with straight lines and full circles for the substrate benzaldehyde and dotted lines and open circles for the product benzoin. The respective parameter estimates are given in Table A.18. The reaction was performed in 50 mM TEA buffer with 2.5 mM MgSO$_4$, 0.5 mM ThDP and 30% DMSO at pH 8.5 and 30°C. The initial concentrations of substrate and enzyme are given in Table A.15.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.38: Fits for the first six minutes of the dimethoxy-benzaldehyde self-ligation (points) and the two-substrate Michaelis-Menten model (model 1, dotted lines). The respective parameter estimates are given in Table A.16 with reaction conditions according to Figure A.36 and Table A.15.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.39: Fits for the first six minutes of the benzaldehyde self-ligation and the two-substrate Michaelis-Menten model (model 1). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.40: Fits for the first six minutes of the dimethoxy-benzaldehyde self-ligation (points) and the macrokinetic model (model 2, dotted line). The respective parameter estimates are given in Table A.16 with reaction conditions according to Figure A.36 and Table A.15.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.41: Fits for the benzaldehyde self-ligation and the macrokinetic model (model 2). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.42: Fits for the dimethoxy-benzaldehyde self-ligation (points) and the microkinetic model (model 3, dotted line). The respective parameter estimates are given in Table A.16 with reaction conditions according to Figure A.36 and Table A.15.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.43: Fits for the benzaldehyde self-ligation and the microkinetic model (model 3). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.44: Fits for the dimethoxy-benzaldehyde self-ligation (points) and the microkinetic model with substrate-dependent enzyme inactivation (model 4, dotted lines). The respective parameter estimates are given in Table A.16 with reaction conditions according to Figure A.36 and Table A.15.
A.5 Progress curve analysis within BioCatNet: comparing kinetic models for enzyme-catalyzed self-ligation

Figure A.45: Fits for the benzaldehyde self-ligation and the microkinetic model with substrate-dependent inactivation (model 4). Symbols as described in Figure A.37. The respective parameter estimates are given in Table A.17 with reaction conditions according to Figure A.37 and Table A.15.
Table A.18: Estimated kinetic parameters for two substrates and the Michaelis-Menten model. Relative standard deviations (%) are given in italics. The residual sum of squares (RSS) over sample size $n$ is indicated for nine experiments containing $n = 2786$ samples for 3,5-dimethoxybenzaldehyde self-ligation (Zavrel et al., 2008) and thirteen experiments containing $n = 374$ samples for benzaldehyde self-ligation (Ohs et al., 2018).

<table>
<thead>
<tr>
<th>Data source</th>
<th>$k_{cat,f}$ [min$^{-1}$]</th>
<th>$K_{m,A}$ [mM]</th>
<th>$\frac{RSS}{n}$ [mM$^2$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,5-dimethoxybenzaldehyde self-ligation</td>
<td>$9.6 \cdot 10^6$</td>
<td>$1.4 \cdot 10^4$</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>$1.8 \cdot 10^4$</td>
<td>$1.8 \cdot 10^4$</td>
<td></td>
</tr>
<tr>
<td>benzaldehyde self-ligation</td>
<td>$5.7 \cdot 10^5$</td>
<td>$3.4 \cdot 10^4$</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>$9.0 \cdot 10^3$</td>
<td>$9.0 \cdot 10^3$</td>
<td></td>
</tr>
</tbody>
</table>